Statistics for Machine Learning

Prof. Kusum Deep
Full Professor (HAG), Department of Mathematics
Joint Faculty, MF School of Data Science and Artificial Intelligence
Indian Institute of Technology Roorkee, Roorkee – 247667

kusum.deep@ma.iitr.ac.in, kusumdeep@gmail.com

Contents

- Introduction to Descriptive Statistics: Mean, Median, Mode, Variance, Correlation.
- Introduction to Probability: Random Variables, Expectation, Probability, Conditional probability, Bayes theorem, law of large numbers
- Introduction to inferential statistics, different types of continuous and discrete probability distributions
- Hypothesis Testing: Sampling Techniques, Central Limit Theorem, Z-test, t-test, F-test. ANOVA , Non-parametric tests: Mann-Whitney, Chi-square test
- Implementation on Python

The Big Picture

Ex: Information about Human beings as a race

Age	**> 0 day**
Food Preference	Vegetarian/ Non- vegetarian/ Vegan
Income per month	>= Rs. 0
Number of friends	Any positive integer
Number of electronic devices they have in their homes	Any positive integer
Own a mobile phone	Yes / No
Number of countries visited	> = 0
Like to watch movies	Yes / No
Witch type of clothes they like to wear	Branded/ unbranded

Statistics and its branches

Statistics is the science of collecting data and analyzing them to infer proportions (sample) that are representative of the population. In other words, statistics is interpreting data in order to make predictions for the population.

There are two branches of Statistics.

1. **DESCRIPTIVE STATISTICS** : Descriptive Statistics is a statistics or a measure that describes the data.
2. **INFERENTIAL STATISTICS** : Using a random sample of data taken from a population to describe and make inferences about the population is called Inferential Statistics.

Descriptive statistics is used to summarize and graph the data for a group and allows you to understand that specific set of observations. It describes a sample. Take a group , record data about the group members, and then use summary statistics and graphs to present the group properties. There is no uncertainty because you are describing only the people or items that you actually measure. You're not trying to infer properties about a larger population.

Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. Because the goal of inferential statistics is to draw conclusions from a sample and generalize them to a population, we need to have confidence that our sample accurately reflects the population. It utilizes sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.

Descriptive statistics

Tools:

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier.

It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Descriptive statistics

Def: Descriptive statistics describe, show, and summarize the basic features of a dataset found in a given study, presented in a summary that describes the data sample and its measurements. It helps analysts to understand the data better.

Randomness and Variability

- Experiments & Processes Are Not Deterministic.
- Statistical techniques are useful for describing and understanding variability.

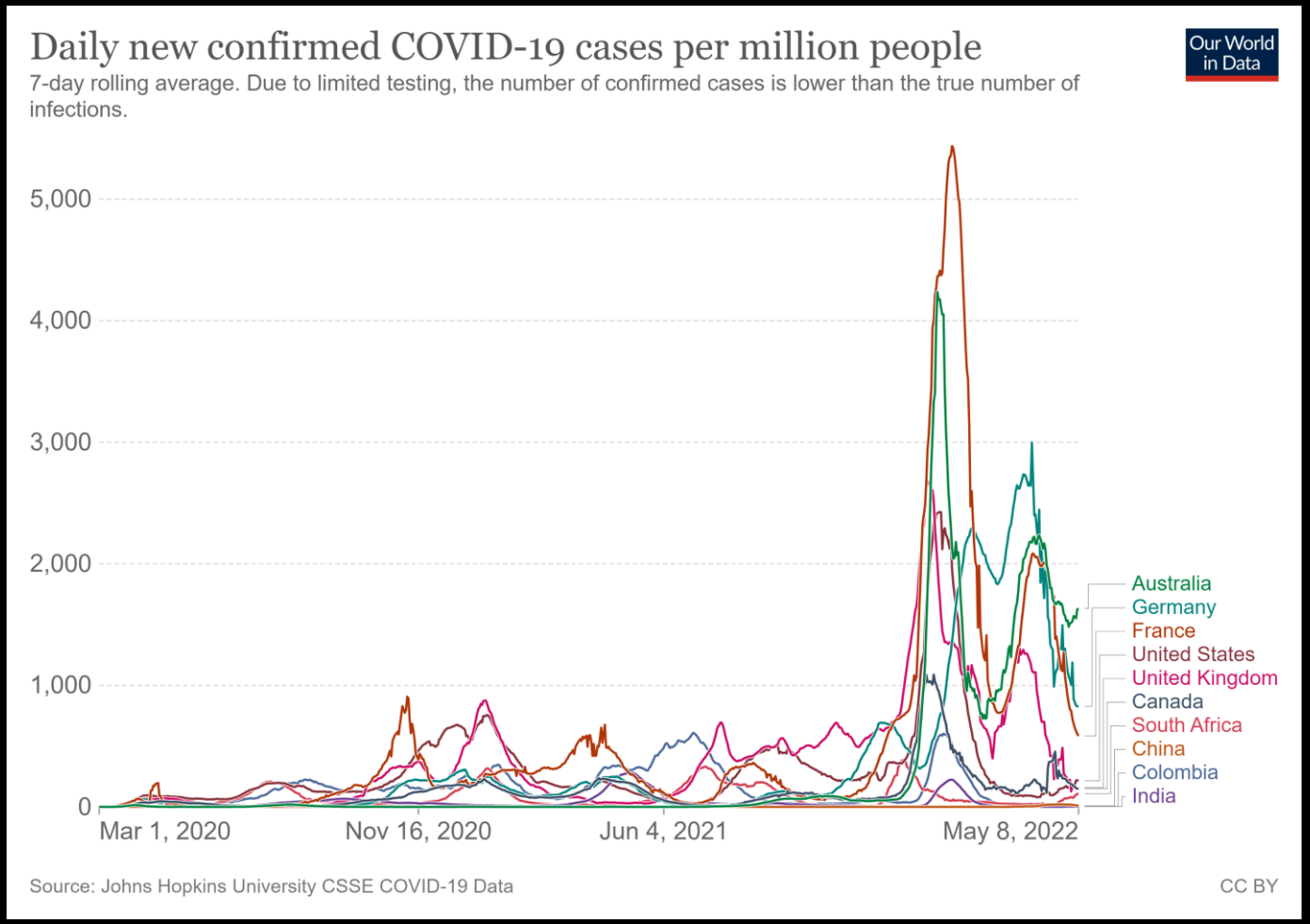
Randomness and Variability

It is the successive observations of a system or phenomenon do ***not*** produce exactly the same result.

E.g. tossing of a coin.

Statistics gives us a framework for describing variability and for learning about potential sources of variability.

Randomness and Variability
Example of variability



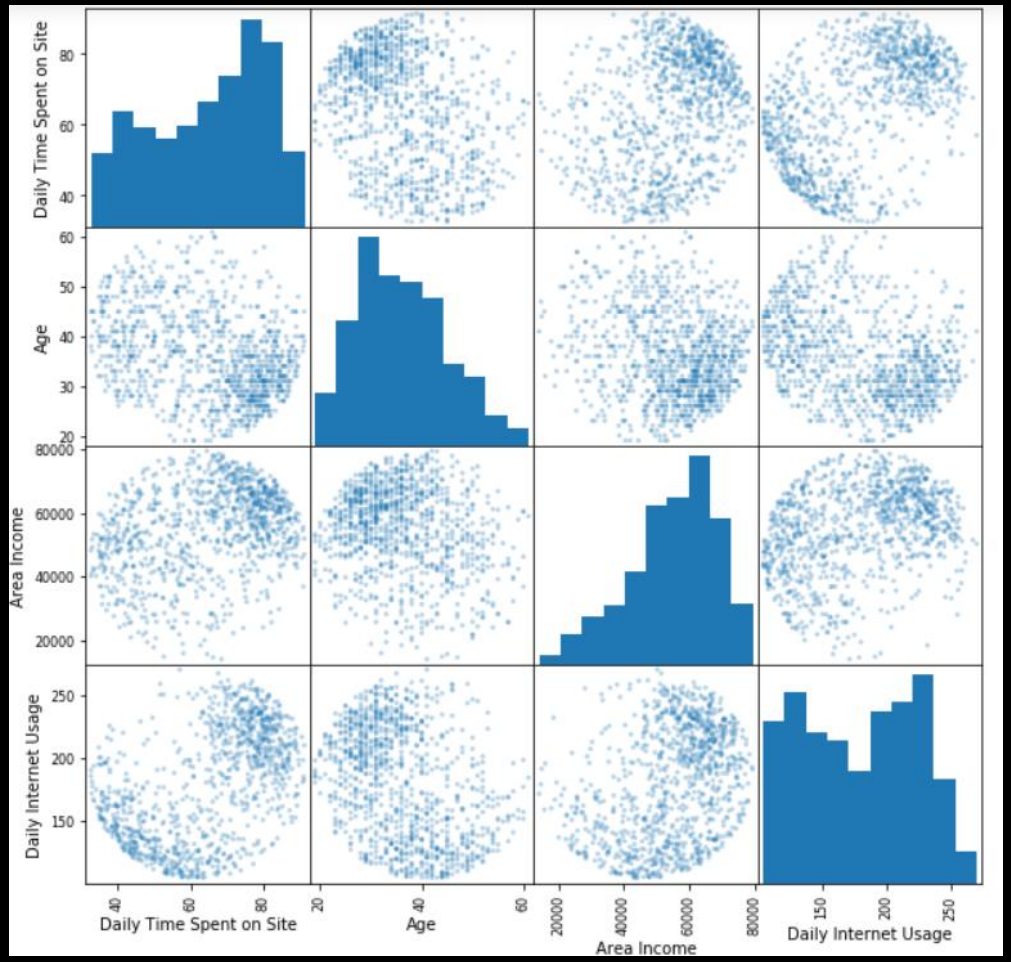
Randomness and Variability Errors

- **Systematic errors,** also called **fixed errors**, are errors associated with using an inaccurate instrument. These errors can be detected and avoided by properly calibrating instruments.

- **Random errors** are generated by a number of unpredictable variations in a given measurement situation.

Randomness often displays an underlying order that can be quantified, and thus used to advantage.

Randomness and Variability
Ex: Predicting Customer Ad Clicks



Fundamental Elements of Statistics

Def: An **experimental unit** is an object about which we collect data.

Person, Place, Disease, Event, etc

- How many persons prefer android phones
- Which places in India are popular tourists spots
- Which disease is occurring more in school children
- How many Gold medals are awarded to Indian team

Fundamental Elements of Statistics

Def: A **population** is a set of units in which we are interested.

Typically, there are too many experimental units in a population to consider *every* one.

Def: If every single data is examined, then it is called a **census**.

Def: A **sample** is a subset of the population.

Fundamental Elements of Statistics

Def: A **variable** is a characteristic or property of an individual unit.

The values of these characteristics will vary.

Ex:

- Highest daily temperature of Delhi.
- No. of monkey pox cases in India.
- Blood pressure of a patient.
- The price of air ticket from Delhi to Mumbai.

Fundamental Elements of Statistics

Number of variables that have been measured:

Def: **Univariate data:** One variable is measured on a single experimental unit.

Def: **Bivariate data:** Two variables are measured on a single experimental unit.

Def: **Multivariate data:** More than two variables are measured on a single experimental unit.

Ex: Give examples of:

1. Univariate data:

   Ex: outcome of tossing a coin

2. Bivariate data:

   Ex: Height and weight of students in a class

3. Multivariate data:

   Ex: Sugar level, BP reading, Temperature, Haemoglobin count

Fundamental Elements of Statistics

Ex: Suppose 10000 green tea consumers are given a blind taste test. Each consumer is asked to state a preference for brand A or brand B.

1. Describe the population.
2. Describe the variable of interest.
3. Describe the sample.
4. Describe the inference.

Fundamental Elements of Statistics

Answers:

1. Because we are interested in the responses of green tea consumers in a taste test, a green tea consumer is the experimental unit. And the population of interest is the collection or set of all green tea consumers.

2. The characteristic that the experimenter wants to measure is the consumer's green tea preference as revealed under the conditions of a blind taste test, so green tea preference is the variable of interest.

Fundamental Elements of Statistics

3. The sample is the 10,000 green tea consumers selected from the population of all green tea consumers.

4. The inference of interest is the ***generalization*** of the green tea preferences of the 10,000 sampled consumers to the population of all green tea consumers. In particular, the preferences of the consumers in the sample can be used to ***estimate*** the percentage of all green tea consumers who prefer each brand.

Fundamental Elements of Statistics

A **measure of reliability** is a statement about the degree of uncertainty associated with a statistical inference.

Based on our analysis, we think 86% of green tea drinkers prefer Brand A to Brand B, ± 4%.

Types of Data

- Qualitative Data
- Quantities Data
  - Discrete
  - Continuous

Types of Data

- **Qualitative Data** are measurements that cannot be recorded on a natural numerical scale, but are recorded in categories.
  - Choice of car brands
  - Side effects of vaccines
  - Choice of elective course
  - Gender
  - Favorite film actor

Types of Data

- **Quantitative Data** are measurements that are recorded on a naturally occurring numerical scale.

Examples:

- Height
- Marks in exam
- Turn over of a company
- Cost of onions

Types of Data

Examples:

- For each mango tree in a grove, the number of ripe mangoes is measured.
  - **Quantitative discrete**

- For a particular day, the number of cars entering IIT Roorkee campus is measured.
  - **Quantitative discrete**

- Time until a light bulb fuses
  - **Quantitative continuous**

Ex: Please type in the chat box the answer to the following:

1. Give 5 examples of qualitative data
2. Give 5 examples of quantitative discrete data
3. Give 5 examples of quantitative continuous data

Sampling Distributions

Sample statistics are used to estimate population parameters.

Def: A **parameter** is a numerical descriptive measure of a population. Its value is almost always unknown.

Def: A **sample statistic** is a numerical descriptive measure of a sample. It can be calculated from the observations.

Sampling Distributions

	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	**σ**	s
Binomial proportion	p	$\hat{p}$

Sampling Distributions

- Numerical descriptive measures calculated from the sample are called **statistics**.
- Since we could draw many different samples from a population, the sample statistic used to estimate the population parameter is itself a **random variable**.

Sampling Distributions

- The **sampling distribution** of a sample statistic calculated from a sample of *n* measurements is the probability distribution of the statistic.
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

Sampling Distributions

Ex: Suppose population = {5, 2, 9, 3, 8)

Draw a sample of size n=3 without replacement and find their average

S1={2, 3, 5} having average = 10/3

S2={5, 9, 3} having average = 17/3

S3={8, 9, 2} having average = 19/3

S4={3, 8, 5) having average = 16/3

Each of the averages are equally likely with probability 1/4

Sampling Distributions

Imagine a very small population consisting of the elements (N=population size=3) 1, 2 and 3. Below are the possible samples (without replacement) of sample size n that could be drawn, along with the means of the samples.

n = 1	$\bar{x}$	n = 2	$\bar{x}$	n = 3 (= N)	$\bar{x}$
1	1	1, 2	1.5		
2	2	1, 3	2	1, 2, 3	2
3	3	2, 3	2.5		

$$\frac{\sum \bar{x}}{3} = 2$$

$$\frac{\sum \bar{x}}{3} = 2$$

$$\frac{\sum \bar{x}}{1} = 2$$

Sampling Distributions

Properties of the Sampling Distribution of $\bar{x}$

The mean of the sampling distribution equals the mean of the population

$$\mu_{\bar{x}} = E(\bar{x}) = \mu$$

The standard deviation of the sampling distribution [the **standard error (of the mean)**)] equals the population standard deviation divided by the square root of *n*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distributions

The Central Limit Theorem

The sampling distribution of $\bar{x}$, based on a random sample of size n observations, will be approximately normal with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma^2}{n}$

The larger the sample size, the better the sampling distribution will approximate the normal distribution.

Ex: Illustrate the Central Limit Theorem with an appropriate example

Point Estimation

The unknown population parameter that we are interested in estimating is called the **target parameter**.

Parameter	Key Word or Phrase	Type of Data
μ	Mean, average	Quantitative
p	Proportion, percentage, fraction, rate	Qualitative
σ	Standard Deviation	Quantitative

Point Estimation

Def: A **point estimator** of a population parameter is a rule or formula that tells us how to use the sample data to calculate a ***single*** number that can be used to ***estimate*** the population parameter.

Point Estimation

We often use the corresponding sample quantity to estimate the population quantity.

Parameter	Statistic	Type of Data
μ	$\bar{X}$	Quantitative
p	$\hat{p}$	Qualitative
σ	s	Quantitative

Point Estimation

Def: If $\hat{\theta}$ is an estimator of θ then the **bias** is defined as $E(\hat{\theta}) - \theta$ and the **variance** is defined as $E\left(\hat{\theta} - E(\hat{\theta})\right)^2$.

The probability distribution considered in the expectation is the sampling distribution of $\hat{\theta}$.

Point Estimation

Ex: Suppose $\theta = \mu$ and $\hat{\theta} = \bar{X}$.

$$E(\hat{\theta}) = E(\bar{X})$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu = \theta.$$

Hence Bias = 0

Such estimators are called **unbiased estimaters**

Point Estimation

Ex: Suppose $\theta = \mu$ and $\hat{\theta} = \bar{X}$.

$$E\left(\hat{\theta} - \theta\right)^2 = E(\bar{X} - \mu)^2$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)\right)^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} E(X_i - \mu)^2$$

The cross terms areeo by independence.

The variance is $\frac{\sigma^2}{n}$

Interval Estimation

Ex: Suppose a sample of 225 college students spend an average of 7 hours per week on social media, with a standard deviation of 3 hours. What can we conclude about the time spent on social media by ALL college students ?

Since the sample size is large, it is not unreasonable to assume that the sample mean is approximately normally distributed.

Interval Estimation

We can be 95% sure that:

$$-1.96 < \frac{\bar{x} - \mu}{\sigma\sqrt{n}} < 1.96$$

That is:

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96$$

Interval Estimation

An **interval estimator** or **confidence interval** is a formula that tell us how to use sample data to calculate an interval that estimates a population parameter.

$$\boldsymbol{\mu = \bar{x} \pm z\sigma_{\bar{x}}}$$

In the college student social media example, the 95% confidence interval is:

$$\left(7 - \frac{1.96 \times 3}{15}, 7 + \frac{1.96 \times 3}{15}\right)$$

$$= (6.608, 7.392)$$

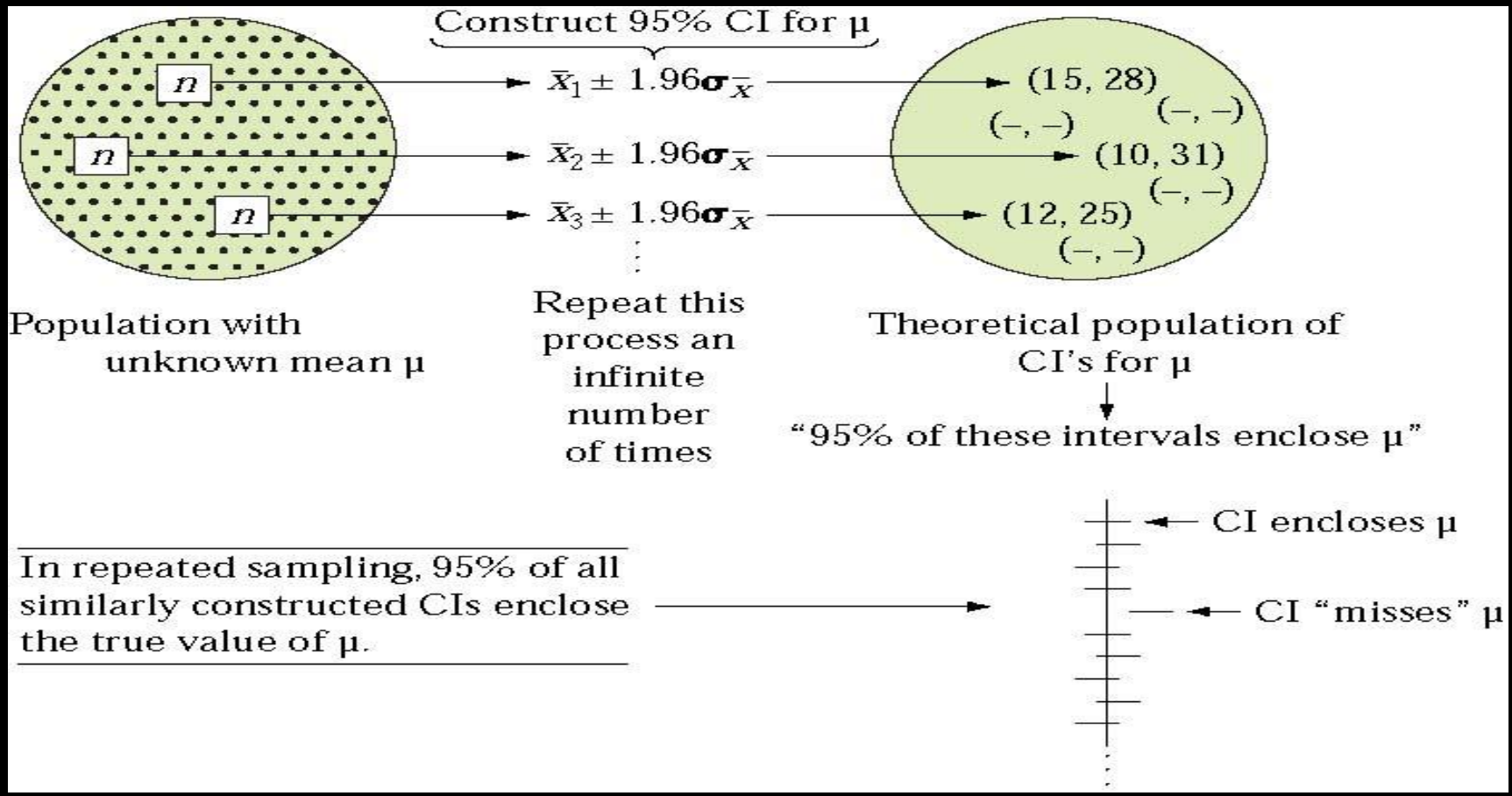
Interval Estimation

- The **confidence coefficient** is the probability that a randomly selected confidence interval encloses the population parameter.
- The **confidence level** is the **confidence coefficient** expressed as a percentage.

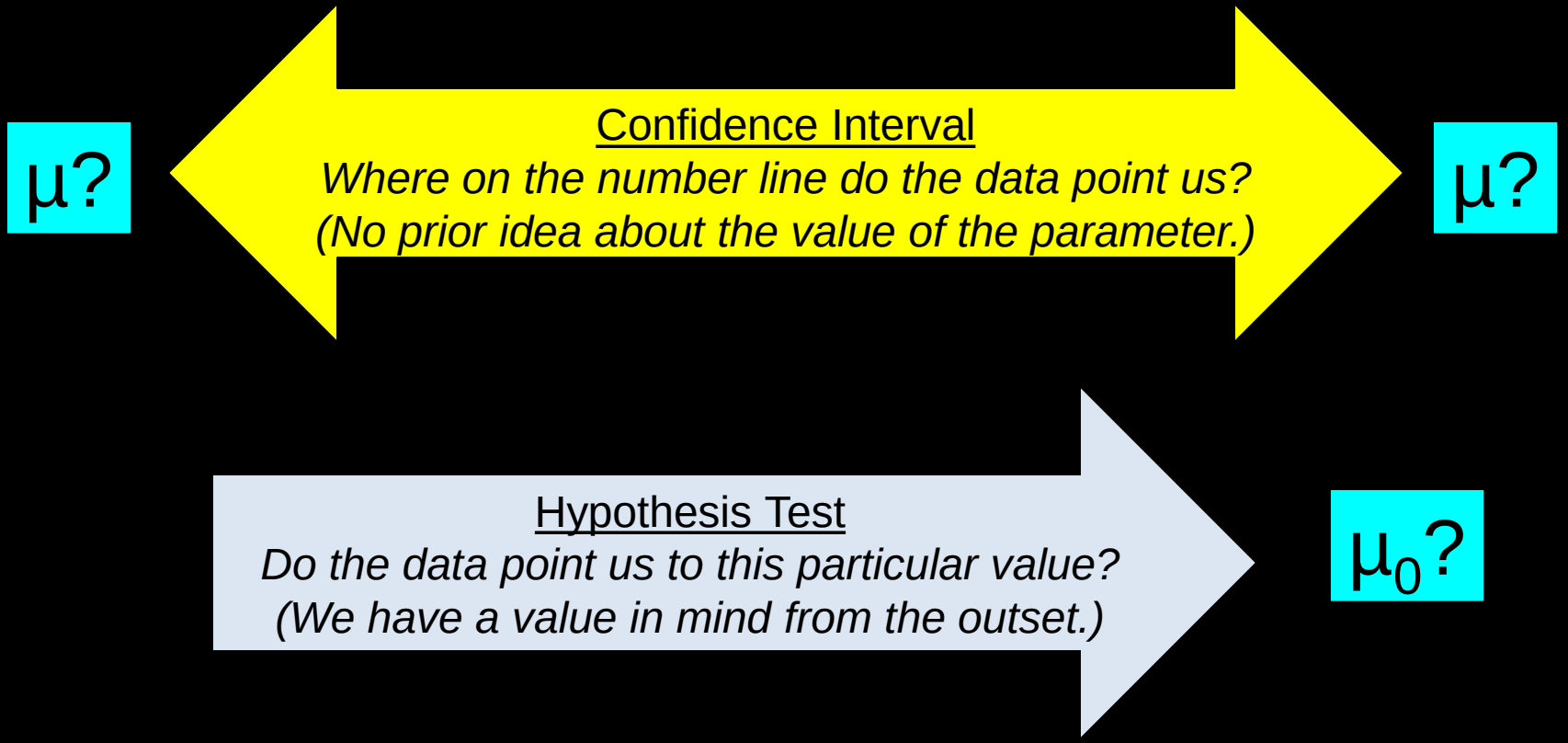
  (Commonly used are 90%, 95% and 99% .

  95% sure $\boldsymbol{\mu = \bar{x} \pm z\sigma_{\bar{x}}}$
- The area outside confidence interval is denoted by α.
- So we are left with (1-95)%=5%=α uncertainty about μ.

Interval Estimation



Hypothesis Testing



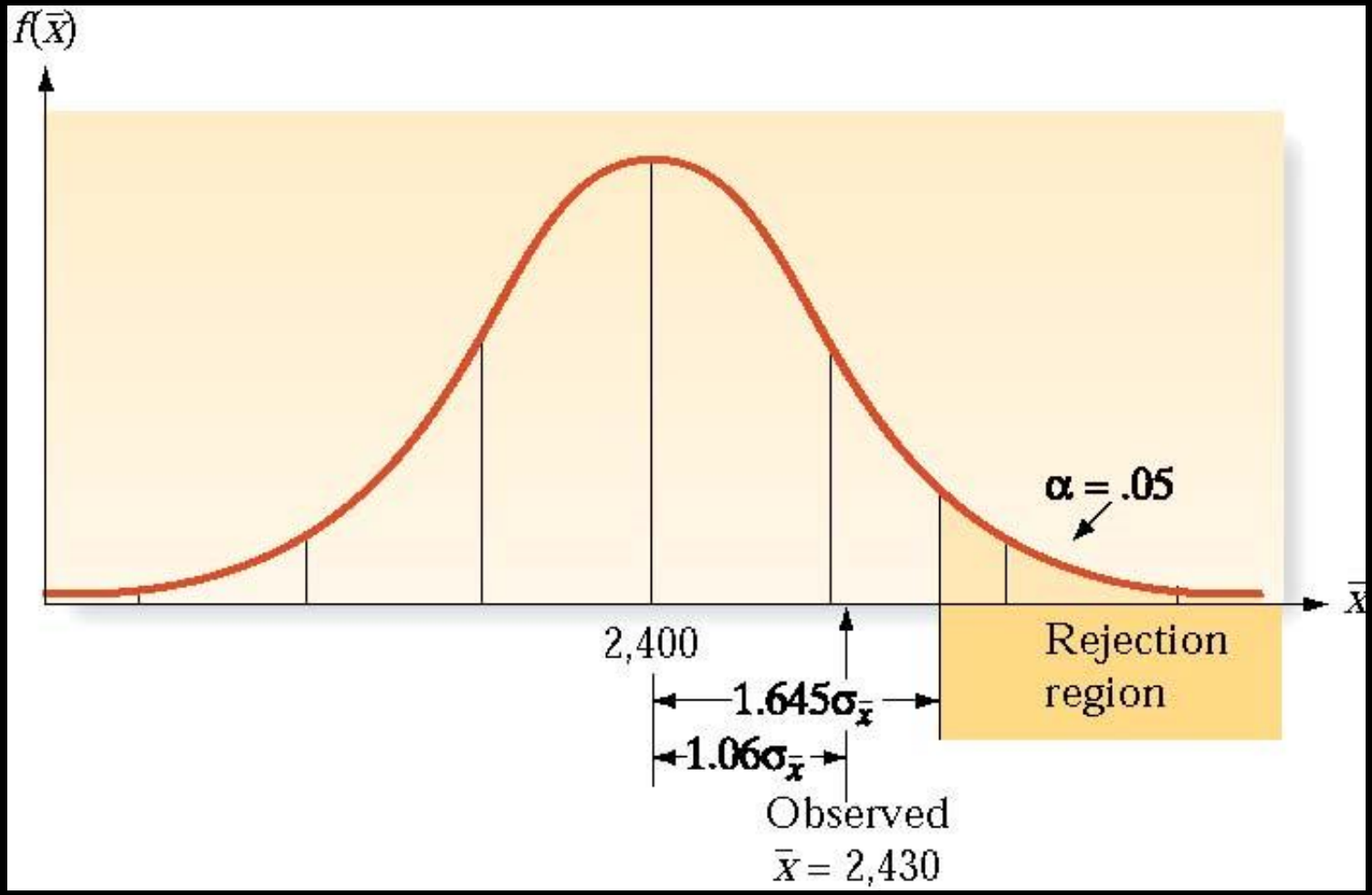
Hypothesis Testing

Null Hypothesis: H_0
- This will be supported unless the data provide evidence that it is false
- The *status quo*

Alternative Hypothesis: H_a
- This will be supported if the data provide sufficient evidence that it is true
- The *research hypothesis*

- If the *test statistic* has a high probability when H_0 is true, then H_0 is *not rejected.*
- If the *test statistic* has a (very) low probability when H_0 is true, then H_0 is *rejected.*

Hypothesis Testing



Hypothesis Testing

Reality ↓ / Test Result →	Do not reject H_0	Reject H_0
H_0 is true	Correct!	Type I Error: rejecting a true null hypothesis P(Type I error) = α
H_0 is false	Type II Error: not rejecting a false null hypothesis P(Type II error) = β	Correct!

Hypothesis Testing

Note: Null hypotheses are either rejected, or else there is insufficient evidence to reject them. (I.e., we don't *accept* null hypotheses.)

Hypothesis Testing

- ***Null hypothesis* (H_0):** A theory about the values of one or more parameters
  - Ex.: H_0: $\mu = \mu_0$ (a specified value for μ)
- ***Alternative hypothesis* (H_a)**: Contradicts the null hypothesis
  - Ex.: H_0: $\mu \neq \mu_0$
- ***Test Statistic*:** The sample statistic to be used to test the hypothesis
- ***Rejection region*:** The values for the test statistic which lead to rejection of the null hypothesis
- ***Assumptions*:** Clear statements about any assumptions concerning the target population
- ***Experiment and calculation of test statistic*:** The appropriate calculation for the test based on the sample data
- ***Conclusion*:** Reject the null hypothesis (with possible Type I error) or do not reject it (with possible Type II error)

Hypothesis Testing

Ex: Suppose a new interpretation of the rules by soccer referees is expected to increase the number of yellow cards per game. The average number of yellow cards per game had been 4. A sample of 121 matches produced an average of 4.7 yellow cards per game, with a standard deviation of 0.5 cards. At the 5% significance level, has there been a change in infractions called?

Hypothesis Testing

H_0: $\mu = 4$ and H_a: $\mu \neq 4$

Sample statistic $\bar{x} = 4.7$

$\alpha = 0.05$

Test statistic: $z* = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{4.7 - 4}{0.045} = 15.56$

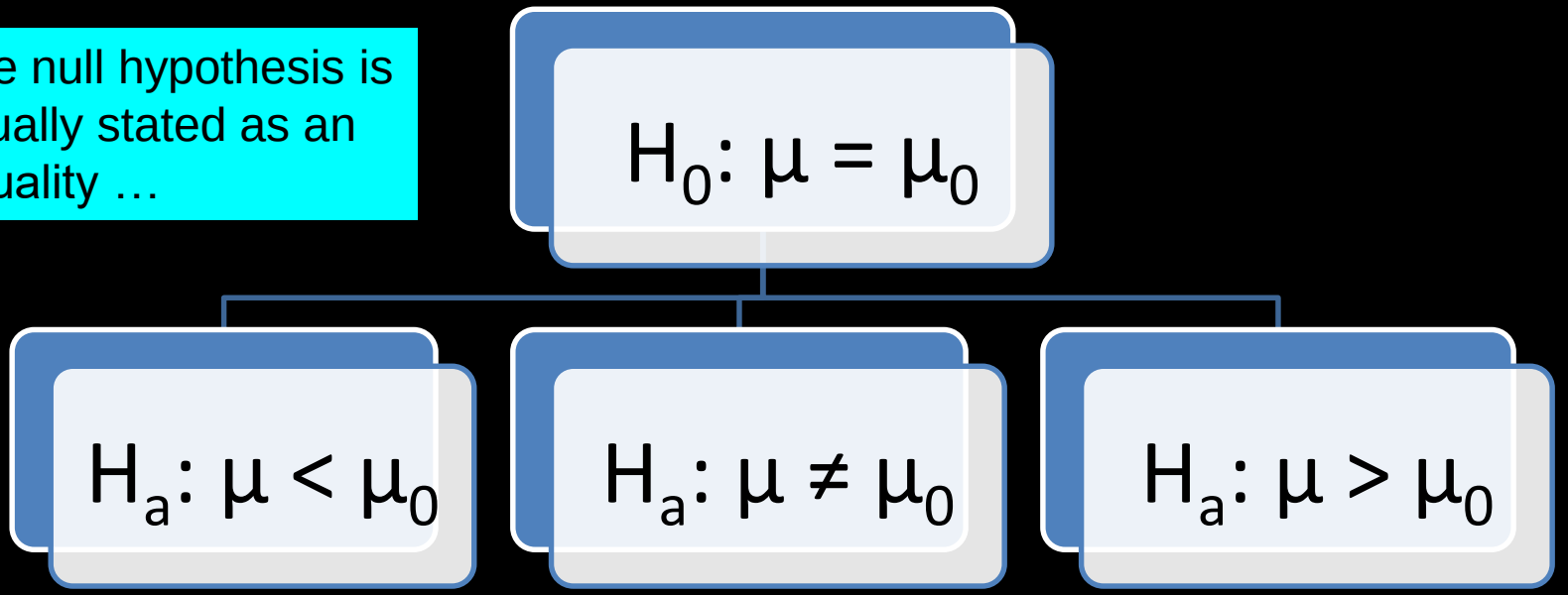
$z_{0.025} = 1.96$

Since $|z*| > 1.96$, we reject the null hypothesis

Conlusion: There has been a change in the infractions called.

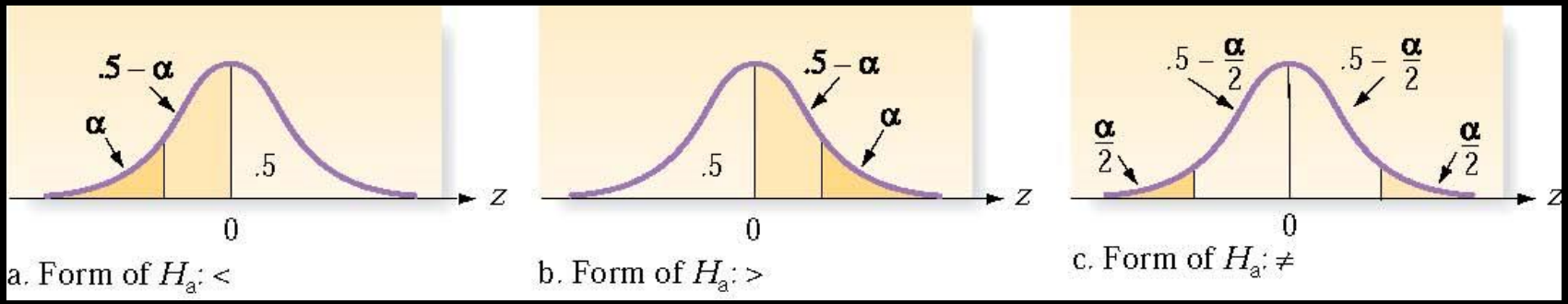
Hypothesis Testing

The null hypothesis is usually stated as an equality …



… even though the alternative hypothesis can be either an equality or an inequality.

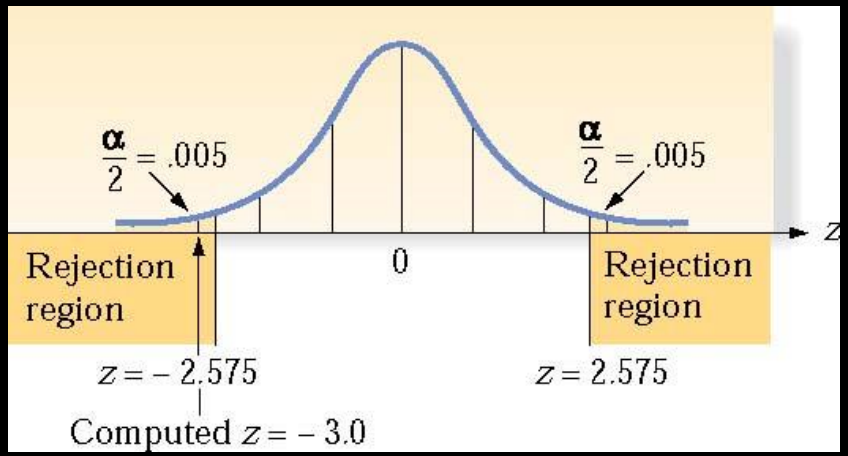
Hypothesis Testing



Hypothesis Testing

Rejection Regions for Common Values of α

	Lower Tailed	Upper Tailed	Two tailed
α = .10	z < - 1.28	z > 1.28	| z | > 1.645
α = .05	z < - 1.645	z > 1.645	| z | > 1.96
α = .01	z < - 2.33	z > 2.33	| z | > 2.575



Large-Sample Test of a Hypothesis about a Population Mean

One-Tailed Test

$$H_0: \mu = \mu_0$$
$$H_a: \mu < or > \mu_0$$

Test Statistic: $z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$

Rejection region: $|z| > z_\alpha$

Two-Tailed Test

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

Test Statistic: $z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$

Rejection region: $|z| > z_{\alpha/2}$

Conditions:

1. A random sample is selected from the target population.
2. The sample size *n* is large.

Hypothesis Testing

Ex: It is reported that a mean salary for males with postgraduate degrees of \$61,340, with an estimated standard error $(s_{\bar{x}})$ equals to \$2,185. We wish to test, at the $\alpha = 0.05$ level, $H_0: \mu = \$60{,}000$.

Please complete……

Hypothesis Testing

Solution:

$$H_0: \mu = 60{,}000$$

$$H_a: \mu \neq 60{,}000$$

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{61340 - 60000}{2185}$$

$$z = 0.613$$

Rejection region is : $|z| > z_{0.025} = 1.96$

Conclusion is Do not reject H_0

Ex: It has been reported that the average credit card debt for college seniors is $3262. The student senate at a large university feels that their seniors have a debt much less than this, so it conducts a study of 50 randomly selected seniors and finds that the average debt is $2995, and the population standard deviation is $1100. Can we support the student senate's claim using the data collected? Conduct the test based on error of $\alpha = 0.05$.

Please complete……

Ans:

Step 1: H_0: $\mu = 3262$ and H_1: $\mu < 3262$

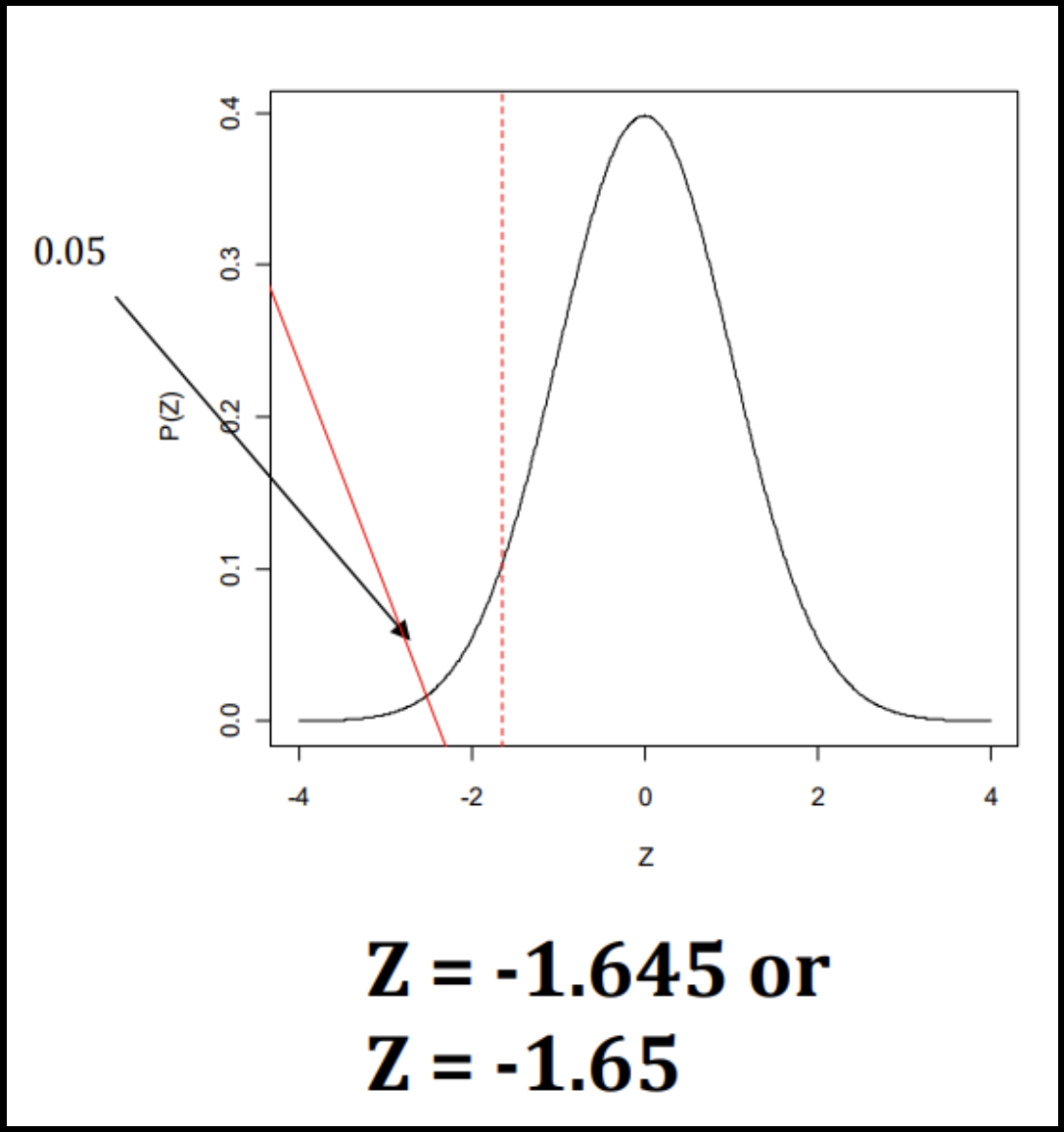
Step 2: Find the critical value(s) from the appropriate table.

Left-tailed test, $\alpha = 0.05, \Rightarrow z$ will be negative and have probability 0.05 underneath it.

Table E The Standard Normal Distribution

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05
−3.4	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0005	.0005	.0005	.0004	.0004	.0004
−3.2	.0007	.0007	.0006	.0006	.0006	.0006
−3.1	.0010	.0009	.0009	.0009	.0008	.0008
−3.0	.0013	.0013	.0013	.0012	.0012	.0011
−2.9	.0019	.0018	.0018	.0017	.0016	.0016
−2.8	.0026	.0025	.0024	.0023	.0023	.0022
−2.7	.0035	.0034	.0033	.0032	.0031	.0030
−2.6	.0047	.0045	.0044	.0043	.0041	.0040
−2.5	.0062	.0060	.0059	.0057	.0055	.0054
−2.4	.0082	.0080	.0078	.0075	.0073	.0071
−2.3	.0107	.0104	.0102	.0099	.0096	.0094
−2.2	.0139	.0136	.0132	.0129	.0125	.0122
−2.1	.0179	.0174	.0170	.0166	.0162	.0158
−2.0	.0228	.0222	.0217	.0212	.0207	.0202
−1.9	.0287	.0281	.0274	.0268	.0262	.0256
−1.8	.0359	.0351	.0344	.0336	.0329	.0322
−1.7	.0446	.0436	.0427	.0418	.0409	.0401
−1.6	.0548	.0537	.0526	.0516	.0505	.0495



Z = -1.645 or

Z = -1.65

Step 4: Compute the test value.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{2995 - 3262}{1100/\sqrt{50}} = -1.716341$$

Step 4: Make the decision to reject or not reject the null hypothesis. Since this is a left-tailed test, our rejection region consists of values of Z that are smaller than our critical value of Z = - 1.645. Since our test value (-1.716341) is less than our critical value -1.645 **we reject the hypothesis.**

Step 5: Summarize the results

We have evidence to support the student senate claim that the university's seniors have credit card debt that is less than the reported average debt.

This is based on a Type I error rate of 0.05.

This means we falsely make the claim above 5% of the time.

Deep Dive into

Descreptive Statistics

The 5 descriptive statistics

There are five most important sample percentiles:

1. The sample minimum (smallest observation)
2. The lower quartile or first quartile.
3. The median (the middle value)
4. The upper quartile or third quartile.
5. The sample maximum (largest observation)

Commonly Used Measures

Def: **Measure of Central Tendency** is a single number summary of the data that typically describes the center of the data. These are of three types:

1. Mean
2. Median
3. Mode

1. **Mean (Average):** Mean is the ratio of the sum of all the observations in the data to the total number of observations.

Mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$

Ex: Let the height of 5 students be:

{150, 155, 153, 151, 157}

Then mean = $\frac{150+155+153+151+157}{5} = 153.2$

2. Median : It is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.

- If the number of observations are odd, median is given by the middle observation in the sorted form.
- If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.

Ex: Let the height of 5 students be:

{150, 155, 153, 151, 157}. Find the median.

Arrange the data in increasing (or decreasing) order:

{150, 151, 153, 155, 157}

Median = 153

Ex 2: Suppose given data is:

{150, 155, 153, 151, 157, 149}. Find the median.

Sorted data is: {149, 150, 151, 153, 155, 157}

Then median = $\frac{151+153}{2} = 152$

3. **Mode**: It is number which has the maximum frequency in the entire data set, or it is the number that appears the maximum number of times. A data can have one or more than one mode.

If the data has one mode it is called **Uni-modal**.

If the data has two modes it called **Bi-modal**.

If the data has more than two modes it is called **Multi-modal**.

Ex: {150, 155, 153, 151, 157} Mode = 1

Ex: {150, 155, 153, 150, 157} Mode = 2

Case I: **Discrete Data Set**

Ex: Calculate the mode for the data pertaining to the size of shoes.

Size of shoes	**4**	**5**	**6**	**7**	**8**
No. of persons	40	55	60	79	50

Mode is 7 as it has the largest frequency.

Case II: **Continuous series data set**

$$Mode = M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times l$$

Where: L = Lower limit of the modal class

l is class interval

f_1 is frequency of modal class

f_0 is frequency of class preceding modal class

f_2 is frequency of class succeeding modal class

Ex: Calculate the mode from the following data related to the marks obtained by the students in a exam.

Marks	No. of students
0-10	3
10-20	4
20-30	6
30-40	17
40-50	23
50-60	22
60-70	20
70-80	10
80-90	5
90-100	2

Modal class is 40-50.

Find the mode. Please complete…..

$$Mode = M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times l$$

$$= 40 + \frac{23 - 17}{2 \times 23 - 17 - 22} \times 10$$

$$= 48.5714$$

Ex: Calculate the mode from the following datausing grouping Method:

X	10	15	20	25	30	35	40
f(x)	9	12	39	34	28	19	9

Grouping Data

x	f(x)	2	3	4	5	6
10	9					
		9+12=21				
15	12		12+39=51			
				9+12+39=60		
20	39				12+39+34=85	
		39+34=73				
25	34					39+34+28=101
			34+28=62			
30	28			34+28+19=81		
		28+19=47				
35	19				28+19+9=56	
			19+9=28			
40	9					

Analysis Table

column no	20	25	30
1	/		
2	/	/	
3		/	/
4		/	/
5	/	/	
6	/	/	/
Total bars	4 bars	5 bars	3 bars

Modal value = 25

Merits of Mode:

1. It can be easily observed from the data.
2. It is easy to compute.
3. It is unaffected by extreme values.
4. Mode can be determined even if the distribution has open end class.
5. It can also be determined easily by graph.
6. It is easy to understand.

Demerits of Mode:

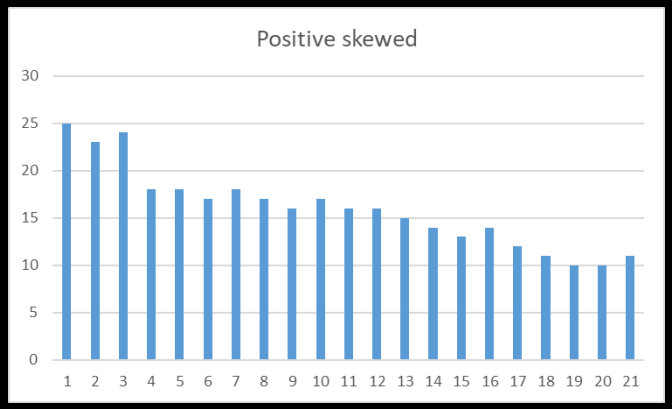
1. Mode is ill-defined when there are distributions with two modes.
2. It is not based on all the values.
3. It cannot be accurate when there are sampling fluctuations.
4. When mode is computed through different methods, the value may differ in each of the methods.

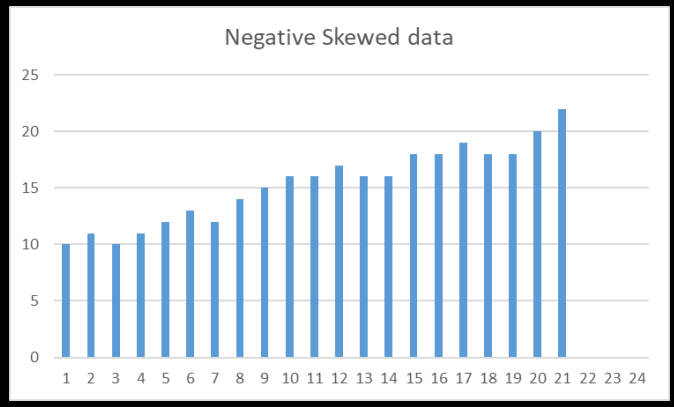
Since mean takes into account “all” data points, it shifts towards the outlier. That is if outlier is big, mean **overestimates** the data and if outlier is small, then the data is **underestimated.**

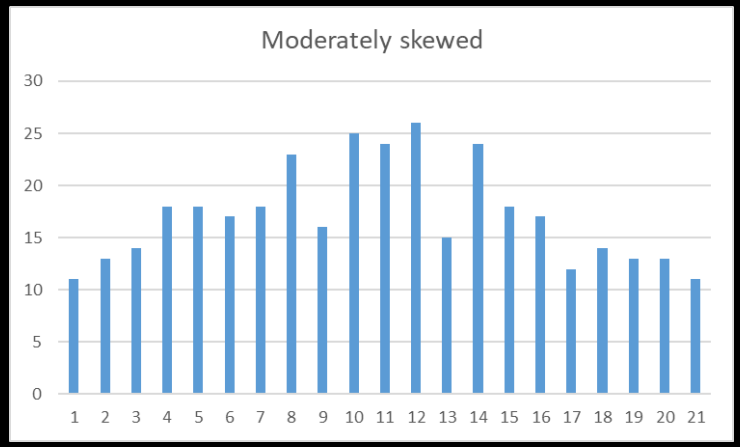
If the distribution is symmetrical, then:

Mean = Median = Mode.

Normal distribution is an example.







	Mean	Median	Mode
Negative skewed	15.333	16	16
Positive skewed	15.9524	16	18
Moderate skewed	17.1429	17	18

Karl Pearson's Formula:

For moderately skewed data:

$$3 \times Median = 2 \times Mean + Mode$$

Ex: $3 \times 17 = 2 \times 17.1429 + 18$

LHS = 51

RHS = 52.2858 ???? Why???

Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)

1. Absolute Deviation from Mean /Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set. It is:

$$= \frac{1}{N} \sum_{i=1}^{N} |x_i - \bar{x}|$$

2. Variance measures how far are data points spread out from the mean.

High variance means data points are spread widely.

Small variance means data points are closer to the mean of the data set. It is:

$$= \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

3. Standard Deviation is the square root of Variance. It is:

$$\sqrt{variance} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

4. Range is the difference between the Maximum value and the Minimum value in the data set. It is:

$$Maximum - Minimum$$

5. Quartiles are the points in the data set that divides the data set into four equal parts.

Q1 is first quartile

Q2 is second quartile

Q3 is third quartile of the data set.

25% of data points lie below Q1 and 75% lie above it.

50% of data points lie below Q2 and 50% lie above it.

Q2 is Median.

75% of data points lie below Q3 and 25% lie above it.

Interquartile range =Q3 – Q1

2. The following data gives the distribution of the marks of 100 students.

Marks	Frequency (f)
Less than 10	5
Less than 20	13
Less than 30	20
Less than 40	32
Less than 50	60
Less than 60	80
Less than 70	90
Less than 80	100

A. Write a Python program to calculate the range and quartiles.

B. Represent above data with help of a plot using Matplotlib.

Please complete……

6. Skewness is the measure of asymmetry in a probability distribution.

It can either be positive, negative or undefined.

Positive Skew: When tail on the right side of the curve is bigger than that on the left side.

$$mean > median > mode$$

Negative Skew: When the tail on the left side of the curve is bigger than that on the right side.

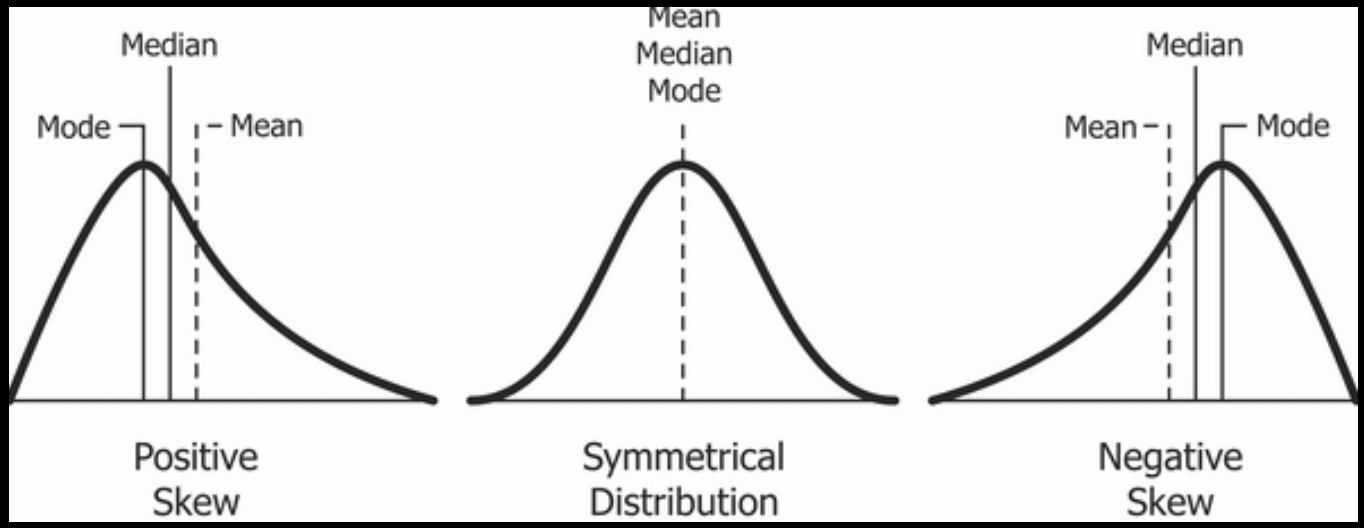
$$mean < median < mode$$

$$skewness = \frac{3(Mean - Median)}{Standard\ Deviation}$$

If it is = 0, the distribution is symmetrical.

If it is < 0, the distribution is Negatively Skewed.

If it is > 0, distribution is Positively Skewed.

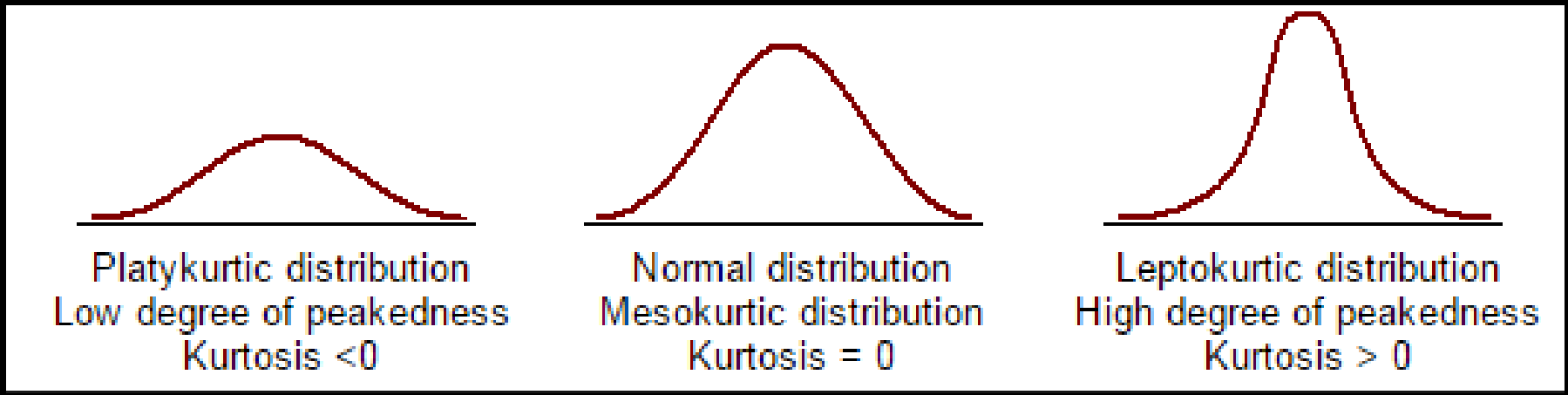


7. Kurtosis describes whether the data is light tailed (lack of outliers) or heavy tailed (outliers present) when compared to a Normal distribution. There are 3 types:

Mesokurtic — This is the case when the kurtosis is zero, similar to the normal distributions.

Leptokurtic — This is when the tail of the distribution is heavy (outlier present) and kurtosis is higher than that of the normal distribution.

Platykurtic — This is when the tail of the distribution is light(no outlier) and kurtosis is lesser than that of the normal distribution.


Platykurtic distribution
Low degree of peakedness
Kurtosis <0
Normal distribution
Mesokurtic distribution
Kurtosis = 0
Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

Ex: The following is the frequency distribution of the number of telephone calls received in 328 successive one minute intervals at an exchange:

Number of Calls	0	1	2	3	4	5	6	7	8	9	10
Frequency	14	21	25	43	51	40	39	12	20	28	35

Write a Python program without using built-in functions to find:

a. Mean
b. Median
c. Mode
d. Range
e. Standard deviation
f. Variance

Using Matplotlib, describe the above visually using a suitable graph/plot.

Please complete…….

8. Correlation: is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).

It is used to test relationships between quantitative variables or categorical variables. The study of how variables are correlated is called **correlation analysis.**

Knowledge of correlations are useful for making predictions. e.g. business, etc…

Examples of data that have a high correlation:

- Your height and your weight.
- Altitude and pressure.
- The amount of time your study and your CGPA.

Examples of data that have a low correlation

- Choice of movies that you prefer and the type of food you like to eat.
- Your pet's name and type of pet food he prefers.
- The cost of a computer and how long it takes to buy a coffee inside a café.

Pearson Correlation Coefficient

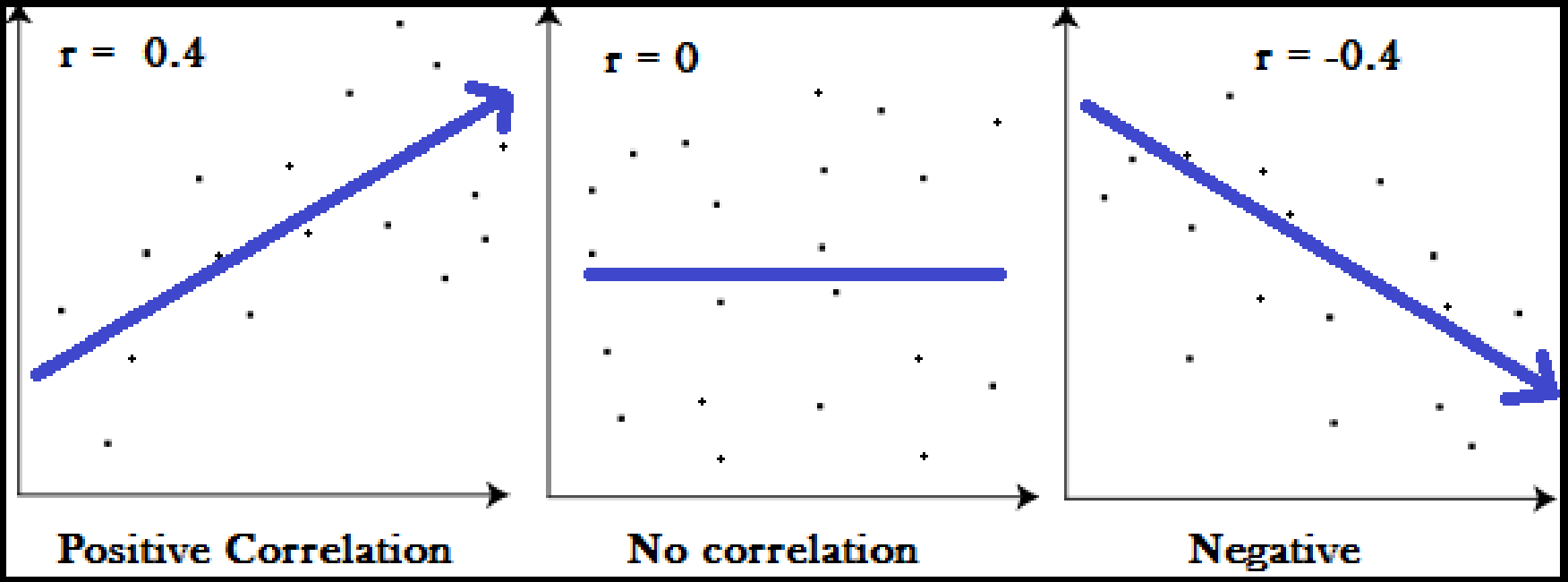
Correlation between: $x_i, i = 1, 2, \ldots n$ and $y_i, i = 1, 2, \ldots n$

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

It lies between -1 and 1.

$$r = \begin{cases} 0\ means\ no\ relationship\ between\ variables \\ -1\ means\ perfect\ negative\ correlation \\ 1\ means\ perfect\ positive\ correlation \end{cases}$$


r = 0.4
r = 0
r = -0.4
Positive Correlation
No correlation
Negative

Ex:

x	y	xy	x^2	y^2
25	11	275	625	121
11	21	231	121	441
82	14	1148	6724	196
77	16	1232	5929	256
49	82	4018	2401	6724
20	69	1380	400	4761
58	73	4234	3364	5329
84	83	6972	7056	6889
55	70	3850	3025	4900
94	66	6204	8836	4356
555	505	29544	38481	33973

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{10 \times 29544 - 555 \times 505}{\sqrt{[10 \times 38481 - 555 \times 555][10 \times 33973 - 505 \times 505]}}$$

$$= 0.188$$

Ex: Write a python code to find correlation from the following data.

Marks	67	97	84	78	73	68	91	87	75
Hours of	20	40	36	34	33	28	42	32	25

Of studying

Please complete…….

Case study
Descriptive Analysis

Overview

In this case study, we are going to investigate product preferences across different types of customers and identify the profile of the typical customer for each TV product offered by Pamsung.

Problem statement

Pamsung is a TV manufacturing company. Its time that they re-invent one of their product lines: the 3D TV, based on the current socio-eco lifestyle. They would like to identify the profile of the typical customer for each TV product.

Requirement

You are from BestMarket, a market research organisation. You are assigned the task to identify the profile of the typical customer for each TV product offered by Pamsung. You are required to investigate whether there are differences across the product lines with respect to customer characteristics.

Data

You may use:

Data set name: Pamsung.csv

Descriptive Analysis

Approach

1. Gathering data
2. Plotting frequency distributions for each attribute of data
3. Plotting box-plots and pair plots
4. Evaluating the statistical parameters
5. Identify product preferences across different types of customers using univariate, bivariate and multi variate analysis
6. Product usage distribution across consumers After understanding the data, perform descriptive analytics to create a customer profile for each Pamsung TV product line

Conclusion

Identified the product preferences across different types of customers and identified the profile of the typical customer for each TV product offered.

Introduction to Probability

- Random variables
- Probability
- Expectation
- Conditional Probability
- Bayes Theorem
- Law of large numbers

Random variables

Def: A **random variable** is a rule that assigns a numerical value to each outcome in a sample space. That is, a random variable is a variable that is subject to random variations so that it can take on multiple different values, each with an associated probability.

Ex: Flipping a coin, rolling a dice, picking a number from a given interval, etc.

They can be of two types: Discrete and continuous

Discrete Random Variable

Ex: Consider the rolling of a six-sided dice.

Its sample space is: {1,2,3,4,5,6}

A random variable X modeling the result of such an experiment would take on a value from the set {1,2,3,4,5,6}, each having probability 1/6.

Continuous Random variable

Ex: Consider the experiment of generating a random number between 0 and 1.

A random variable modeling the result of such an experiment could take on any real number in the interval [0,1], where each number would be equally likely.

Expected value

The **mean**, **expected value**, or **expectation** of a random variable X is written as E(X) or μ_x.

If we observe N random values of X, then the mean of the N values will be approximately equal to E(X) for large N.

Def: Let X be a continuous random variable with p.d.f. $f_X(x)$. Then the expected value of X is

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

Def: Let X be a discrete random variable with probability function $f_X(x)$.

Then, the expected value of X is

$$E(X) = \sum_x x f_X(x) = \sum_x x Pr(X = x)$$

Def: Let X be a continuous random variable and let $g(X)$ be a function. Then expected value of $g(X)$ is:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Def: Let X be a discrete random variable and let $g(X)$ be a function. Then, the expected value of $g(X)$ is:

$$E(g(X)) = \sum_x \ g(x) f_X(x) = \sum_x \ g(x) Pr(X = x)$$

Expected value of X and Y

If X and Y are two random variables then they may be independent or may be dependent.

Let $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$ be a large number of pairs of observations, then as N tends to infinity, the average $\frac{1}{N}\sum_{i=1}^{N} x_i y_i$ approaches the expectation E(XY).

Properties of expectation

1. Let g and h be functions, and let a and b be constants. For any random variable X (discrete or continuous),

   $$E\{ag(X) + bh(X)\} = aE\{g(X)\} + bE\{h(X)\}$$

   In particular, $E\{aX + b\} = aE(X) + b$

2. Let X and Y be any random variables (discrete/ continuous/ independent/non-independent) , then:

   $$E(X + Y) = E(X) + E(Y)$$

3. Let X and Y be independent random variables, and g,h are functions. Then :

$$E(XY) = E(X)E(Y)$$

$$E\big(g(X)h(Y)\big) = E\big(g(X)\big)E(h(Y))$$

4. $E(XY) = E(X)E(Y)$ is only generally true if X and Y are INDEPENDENT.

5. If X and Y are independent then $E(XY) = E(X)E(Y)$. However converse is not generally true.

Probability

Def: Ratio between number of favourable outcomes and total number of outcomes

Ex: Throw a die once. Random Variable X = "The score shown on top face". X could be 1, 2, 3, 4, 5 or 6.

Sample Space is {1, 2, 3, 4, 5, 6}

P(X = 1) = 1/6 P(X = 2) = 1/6 P(X = 3) = 1/6

P(X = 4) = 1/6 P(X = 5) = 1/6 P(X = 6) = 1/6

sum of the probabilities = **1**

Ex: How many heads will come when we toss 3 coins?

Let X = "The number of Heads" be the Random Variable. Possibilities are: 0 Heads, 1 Head, 2 Heads or 3 Heads. So the Sample Space = {0, 1, 2, 3}

But this time the outcomes are NOT all equally likely. The three coins can land in eight possible ways: HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.

$P(X = 3) = 1/8$

$P(X = 2) = 3/8$

$P(X = 1) = 3/8$

$P(X = 0) = 1/8$

Ex: Two dice are tossed. Random Variable is X = "The sum of the scores on the two dice".

All possible values are:

		First die					
Second die		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Sample = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

There are 36 possibilities.

2 occurs just once, so $P(X = 2) = 1/36$

3 occurs twice, so $P(X = 3) = 2/36 = 1/18$

4 occurs three times, so $P(X = 4) = 3/36 = 1/12$

5 occurs four times, so $P(X = 5) = 4/36 = 1/9$

6 occurs five times, so $P(X = 6) = 5/36$

7 occurs six times, so $P(X = 7) = 6/36 = 1/6$

8 occurs five times, so $P(X = 8) = 5/36$

9 occurs four times, so $P(X = 9) = 4/36 = 1/9$

10 occurs three times, so $P(X = 10) = 3/36 = 1/12$

11 occurs twice, so $P(X = 11) = 2/36 = 1/18$

12 occurs just once, so $P(X = 12) = 1/36$

Case 1: What is the probability that the sum of the scores is 5, 6, 7 or 8?

In other words: What is $P(5 \leq X \leq 8)$?

$P(5 \leq X \leq 8) = P(X=5) + P(X=6) + P(X=7) + P(X=8)$

$= (4+5+6+5)/36 = 5/9$

Case 2: If $P(X=x) = 1/12$, what is the value of x?

$P(X=4) = 1/12$, and

$P(X=10) = 1/12$

So there are two solutions: $x = 4$ or $x = 10$

Question 1: Three horses A, B and C are in a race; A is twice as likely to win as 3 and B is twice as likely to win as C. What are their respective probabilities of winning, i.e. P(A),P(B)and P(C)?

Solution: let P(C) = p,

Then, according to problem P(B) = 2p and P(A) = 2P(B) = 4p

Since, sum of probabilities must be 1; hence

4p+2p+p = 1 => p = 1/7

Thus, P(A) = 4/7, P(B) = 2/7 and P(C) = 1/7.

Question 2: What is the probability that B or C wins, i.e. P({B,C})?

Solution: P({B,C}) = P(B) + P(C) = 3/7

Question 3: Let a die be weighted so that the probability of a number appearing when the die is tossed is proportional to the given number (e.g. 6 has twice the probability of appearing as 3). Let A = {even number}, B = {prime number}, C = (odd number).

(i) Describe the probability space, i.e. find the probability of each sample point.

(ii) Find P(A),P(B)and P(C).

(iii) Find the probability that: (a) an even or prime number occurs; (b) an odd prime number occurs; (c)A but not B occurs.

Solution:

Let P(1) = p. Then P(2)= 2p, P(3) = 3p, P(4) = 4p, P(5)= 5p and P(6) = 6p. Since the sum of the probabilities must be one, we obtain p + 2p + 3p + 4p+ 5p + 6p = 1 or p = 1/21. Thus P(1) = 1/21, P(2)= 2/21, P(3) = 1/7, P(4) = 4/21, P(5)=5/21 &, P(6)= 2/7.

P(A) = P({2,4,6}) = 4/7, P(B) = P({2,3,5}) = 10/21, P(C) = P({1,3,5}) = 3/7.

(a) The event that an even or prime number occurs is A UB = {2,4,6,3,5}, or that 1 does 20not occur. Thus P(A UB) = 1 -P(1) = 20/21.

(b) The event that an odd prime number occurs is B∩C = {3,5}. Thus P(B∩C) = P({3,5}) = 8/21.

(c) The event that A but not B occurs is A∩ B^c = {4,6}. Hence P(A∩ B^c) = P({4,6}) = 10/21.

The probability density function

Let I_x is a small interval around x.

Then, assuming ρ is continuous, the probability that X is in that interval will depend both on the density ρ(x) and the length of the interval:

$$\Pr(X \in I_x) \approx \rho(x) \times length\ of\ I_x$$

$$\Pr(x \in A) = \int_A \rho(x)dx$$

If $I = [a, b]\ where\ a \leq b$, then the probability that $a \leq X \leq b$ is:

$$\Pr(x \in I) = \int_I \rho(x)dx = \int_a^b \rho(x)dx$$

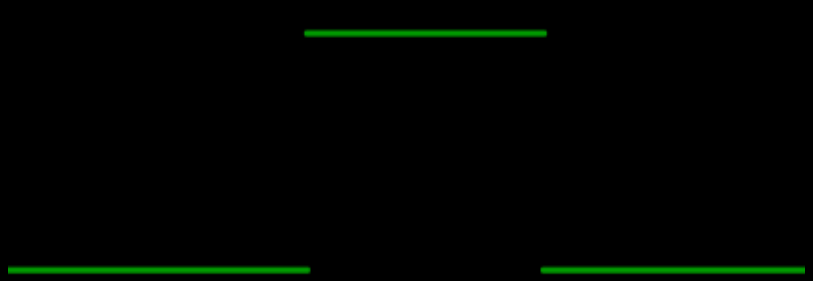
For probability density function,

$\rho(x) \geq 0$ for all x

And $\int \rho(x)dx = 1$

Ex: Let X be given by a uniform distribution in the interval [0,1]. The probability density function of X is given by:

$$\rho(x) = \begin{cases} 1 & if\ x \in [0,1] \\ 0 & otherwise \end{cases}$$



If I is an interval contained in [0,1], say $I = [a, b]$, $such\ that\ 0 \leq a \leq b \leq 1$, then $\rho(x)$=1 in the interval and

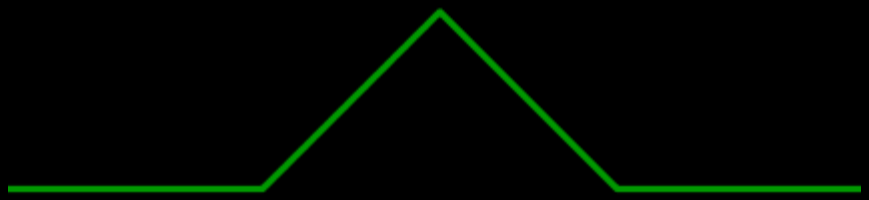
$$\Pr(x \in I) = \int_I \rho(x)dx$$

$$= b - a \quad \text{(length of the interval)}$$

For any interval I, $\Pr(x \in I)$ = length of the intersection of I with the interval [0,1].

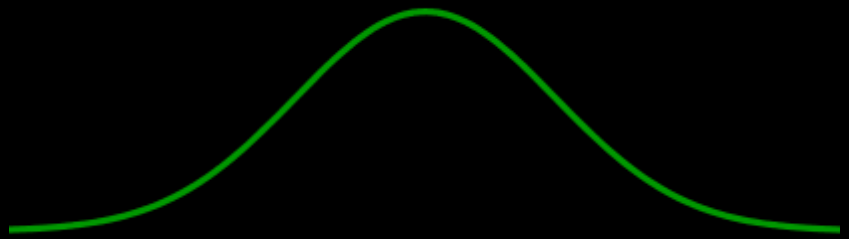
Ex 2: If $\rho(x) = \begin{cases} x \; if \; 0 < x < 1 \\ 2 - x \; if \; 1 < x < 2 \\ 0 \; otherwise \end{cases}$

Then $\rho(x)$ is a triangular probability density function centered around 1.



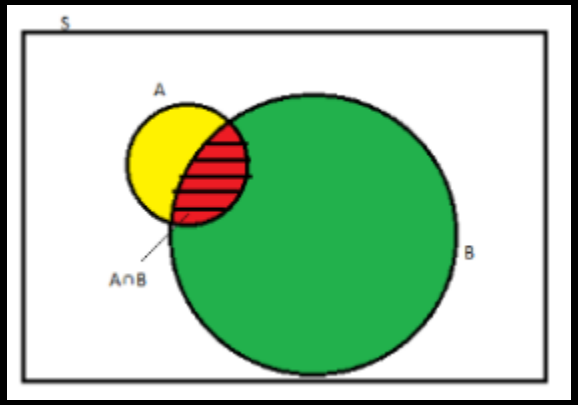
Ex 3: Probability density function of a Gaussian random variable, also called a normal random variable. It looks like a bell-shaped curve.

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



Conditional Probability

Def: The probability of occurrence of any event A when another event B in relation to A has already occurred is known as conditional probability. It is denoted by P(A|B)



Sample space is S

A and B are two events.

When an event B has already occurred, then sample space S gets reduced to B.

Independent Events: Each event is not affected by any other events.

Ex: Tossing a coin.

Each toss is a perfectly isolated. What was the outcome in the past will not affect the current toss.

Probability of head = ½.

Dependent Events: If the occurrence of an event can be affected by previous event.

Ex: Let there be 2 blue and 3 red balls in a bag. What are the chances of drawing a blue ball?

Chances of blue ball in 1^{st} pick = 2/5.

Next time chances will change.

If 1^{st} pick was red, then chance of blue ball in 2^{nd} pick = 2/4

If 1^{st} pick was blue, then chances of blue ball in 2^{nd} pick = 1/4

This is because balls are removed from the bag.

So the next event depends on what happened in the previous event.

If the balls are replaced back in the bag, every time, then the chances do not change and the events are **independent**.

Conditional Probability of A given B

$= P(A|B) = P(A \cap B)/P(B)$

Ex: The probability that it is Friday and that a student is absent is 0.03. Since there are 5 school days in a week, the probability that it is Friday is 0.2. What is the probability that a student is absent given that today is Friday?

Sol: Conditional probability is: $P(B|A) = P(A \cap B)/P(A)$

$P(\text{Absent} \mid \text{Friday}) = P(\text{Absent and Friday})/P(\text{Friday})$

$= 0.03/0.2$

$= 0.15$

Ex 2: A teacher gave her students two tests in maths and English. 25% of the students passed both the tests and 40% of the students passed the maths test. What percent of those who passed the maths test also passed the English test?

Sol:

Let A be the event of the no. of students who passed maths tests.

Let B be the event of the no. of students who passed maths and English tests.

$P(A) = 40\% = 0.40$

$P(A \cap B) = 25\% = 0.25$

Percent of students who passed the maths test also passed the English test

= Condition probability of B given A

$= P(B|A) = P(A \cap B)/P(A)$

$= 0.25/0.40$

$= 0.625$

Ex 3: A bag contains green and yellow balls. Two balls are drawn without replacement. The probability of selecting a green ball and then a yellow ball is 0.28. The probability of selecting a green ball on the first draw is 0.5. Find the probability of selecting a yellow ball on the second draw, given that the first ball drawn was green.

Sol: Let A and B be the events of drawing a green ball in the first draw and yellow ball in the second draw, respectively.

Then, $P(A) = 0.5$ and $P(A \cap B) = 0.28$

Probability of selecting a yellow ball on the second draw, given that the first ball drawn was green = Conditional of B given A

$= P(B|A) = P(A \cap B)/P(A)$

$= 0.28/0.5$

$= 0.56$

Ex: Let a pair of fair dice be tossed. If the sum is 6, find the probability that one of the dice is a 2. In other words, if E = {sum is 6} = {(1,5), (2,4), (3,3), (4,2), (5,1)}

and A = {a 2 appears on at least one die} find P(A|E)

Solution Now E consists of five elements and two of them, (2,4) and (4,2), belong to A: $A \cap E$ = {(2,4), (4,2)}. Then P(A| E) = 2/5.

On the other hand, since A consists of eleven elements, =A ((2, 1), (2, 2),(2,3), (2,4), (2,5), (2,6),(1,2), (3, 2),(4,2), (5,2), (6, 2)) and S consists of 36 elements, P(A) = 11/36.

Ex: Three fair coins are tossed. Find the probability p that they are all heads if the first coin is heads

one of the coins is heads.

Solution: The sample space has eight elements: S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

If the first coin is heads, the reduced sample space is A = {HHH,HHT,HTH,HTT}. Since the coins are all heads in 1 of 4 cases, p = 1/4.

If one of the coins is heads, the reduced sample space is B = {HHH, HHT, HTH, HTT, THH, THT, TTH}. Since the coins are all heads in 1 of 7 cases, p = 1/7.

Ex: Two digits are selected at random from the digits 1 through 9. If the sum is even, find the probability p that both numbers are odd.

Solution: The sum is even if both numbers are even or if both numbers are odd. There are 4 even numbers (2,4,6,8); hence there are $C_2^4 = 6$ ways to choose two even numbers. There are 5 odd numbers (1,3,5,7,9); hence there are $C_2^5 = 10$ ways to choose two odd numbers. Thus there are 6 + 10 = 16 ways to choose two numbers such that their sum is even; since 10 of these ways occur when both numbers are odd, p = 10/16 = 5/8.

Bayes Theorem

If A and B are events, such that $P(B) \neq 0$, then

$P(A|B) = P(B|A)\ P(A)/P(B)$

Where $P(A|B)$ is conditional probability that the probability of event A occurring given that B is true.

And $P(B|A)$ is conditional probability that the probability of event B occurring given that A is true.

And P(A) and P(B) are probabilities of observing A and B, respectively without any given conditions.

Bayes' rule and computing has applications in solving puzzles like:

- Three Prisoners problem,
- Monty Hall problem,
- Two Child problem and
- Two Envelopes problem.

Ex: Cancer rate

Even if 100% of patients with pancreatic cancer have a certain symptom, when someone has the same symptom, it does not mean that this person has a 100% chance of getting pancreatic cancer.

Assuming the incidence rate of pancreatic cancer is 1/100000, while 10/99999 healthy individuals have the same symptoms worldwide, the probability of having pancreatic cancer given the symptoms is only 9.1%, and the other 90.9% could be "false positives" (that is, falsely said to have cancer; "positive" is a confusing term when, as here, the test gives bad news).

Based on incidence rate, the following table presents the corresponding numbers per 100,000 people.

Cancer Symptoms	Yes	No	Total
Yes	1	0	1
No	10	99989	99999
Total	11	99989	100000

Probability of having cancer when you have the symptoms:

$$P(Cancer|Symptoms) = \frac{P((Symptoms|Cancer)P(Cancer)}{P(Symptoms)}$$

$$= \frac{P((Symptoms|Cancer)P(Cancer)}{P(Symptoms|Cancer)P(Cancer) + P(Symptoma|noncancer)P(noncancer)}$$

$$= \frac{1 \times 0.00001}{1 \times 0.00001 + (10/99999 \times 0.99999}$$

$$= \frac{1}{11} \approx 0.1\%$$

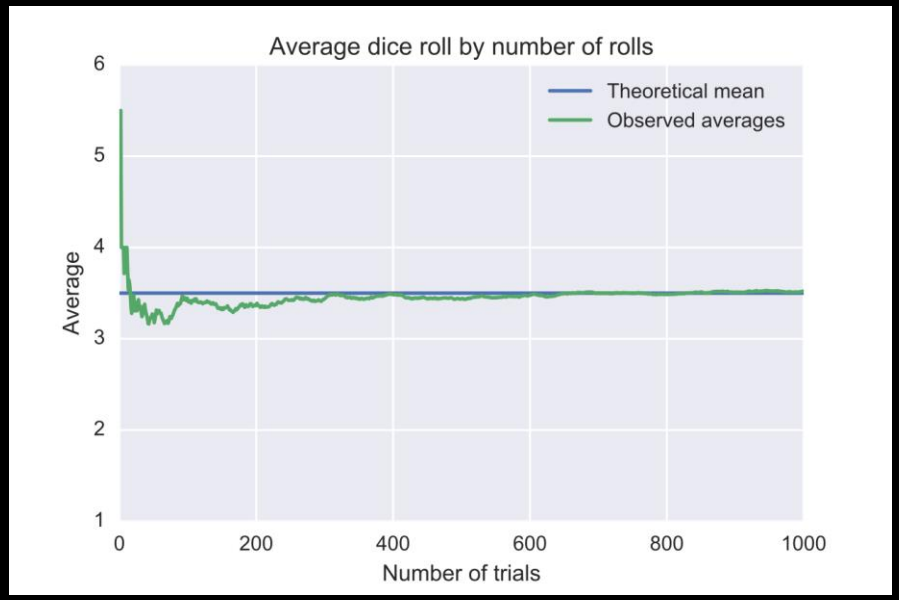
Law of Large numbers

Law of large numbers describes the result of performing the same experiment a large number of times. The *average* of the results obtained from a large number of trials should be close to the *expected value* and tends to become closer to the expected value as more trials are performed.

$$\lim_{n\to\infty}\left(\sum_{i=1}^{n}\frac{x_i}{n}\right) = \bar{x}$$

Ex: Single roll of a fair, six-sided dice produces one of the numbers 1, 2, 3, 4, 5, or 6, each with equal probability. Therefore, the expected value of the average of the rolls is: $= \frac{1+2+3+4+5+6}{6} = 3.5$

If a large number of six-sided dice are rolled, the average of their values will approach 3.5, with the precision increasing as more dice are rolled.



Ex: **Monte Carlo methods** are a class of computational algorithms that rely on repeated random sampling to obtain numerical results. The larger the number of repetitions, the better the approximation tends to be. The reason that this method is important is that it is difficult or impossible to use other approaches.

Ex: In the business and finance, Law of Large Numbers is related to growth rates of businesses.

It states that as a company grows, it becomes more difficult to sustain its previous growth rates. Thus, the company's growth rate declines as it continues to expand. The law of large numbers may consider different financial metrics such as market capitalization, revenue, and net income.

Ex: Company ABC's market capitalization is $1 million while Company XYZ's market capitalization is $100 million. Company ABC experiences a significant growth of 50% per year. For ABC, the growth rate is easily attainable since its market capitalization only grows by $500,000.

For Company XYZ, that growth rate is almost impossible because it implies that its market capitalization should grow by $50 million per year. Note that the growth of Company ABC will decline over time as it continues to expand.

Ex: Insurance companies also rely on the law of large numbers to remain profitable.

They can provide insurance to thousands of individuals who pay a certain premium each month and only a small percentage of the individuals they ensure will actually need to use the insurance to pay for large unexpected expenses.

For example, 1,000 people might each pay $1,000 per year for insurance, which results in a profit of $1,000,000 for an insurance company.

However, 90 people might each need to receive $10,000 from the insurance company to cover unexpected expenses from various accidents, which results in a $900,000 loss for the insurance company.

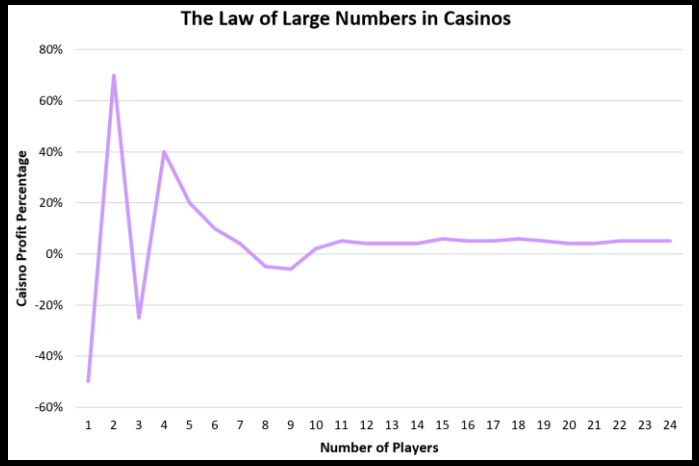
In the end, the insurance company earns a profit of $1,000,000 – $900,000 = $100,000.

This means that the insurance company can expect to earn a fairly predictable profit, on average, across thousands of individuals.

Note that this business model works because an insurance company ensures a **large number of people**. If they only ensured 10 people it would be far too risky because a large unexpected expense could wipe out the business.

The Law of Large Numbers in Casinos

Casinos rely on the law of large numbers to reliably produce profits. For most games, the casino wins about 51-55% of the time. This means that individuals can get lucky and win a decent amount from time to time, but over the course of tens of thousands of individual players, the casino will win the expected 51-55% of the time.



Ex: Jessica might play a few games at the casino and win $50.

Mike might play a few games as well and lose $70.

John might play a few games and win $25.

Susan might play a few games and lose $40.

Some players will win money and some will lose money, but because of the way the games are designed the casinos can be sure that they'll win over the course of thousands of individuals.

Probability distribution

Def: A probability distribution is a function that describes how likely you will obtain the different possible values of the random variable.

Ex: Suppose two dice are rolled and the sum of the upper face is recorded. Likely outcomes or sample space will be: {2,3,4,5,6,7,8,9,10,11,12}

It is more likely that the outcome will be 7 or 8 as compared to 2 and 12.

Discrete probability distribution

The sample space could be :

1. Finite, e.g. flipping a coin: {head, tail}
2. Countably infinite, e.g.: {0,1,−1,2,−2,3,−3,…}.

For a discrete random variable X, we form its probability distribution function by assigning a probability that X is equal to each of its possible values.

Discrete probability distribution

Ex: For a dice, each of the outcome has a probability of 1/6.

Probability distribution function is a probability mass function.

Def: The probability mass function P(x) for a random variable X is defined so that for any number x, the value of P(x) is the probability that the random variable X equals the given number x, i.e., P(x)=Pr(X=x).

$P(x) \geq 0 \;\; for\; all\; x$

$\sum_x \; P(x) = 1$

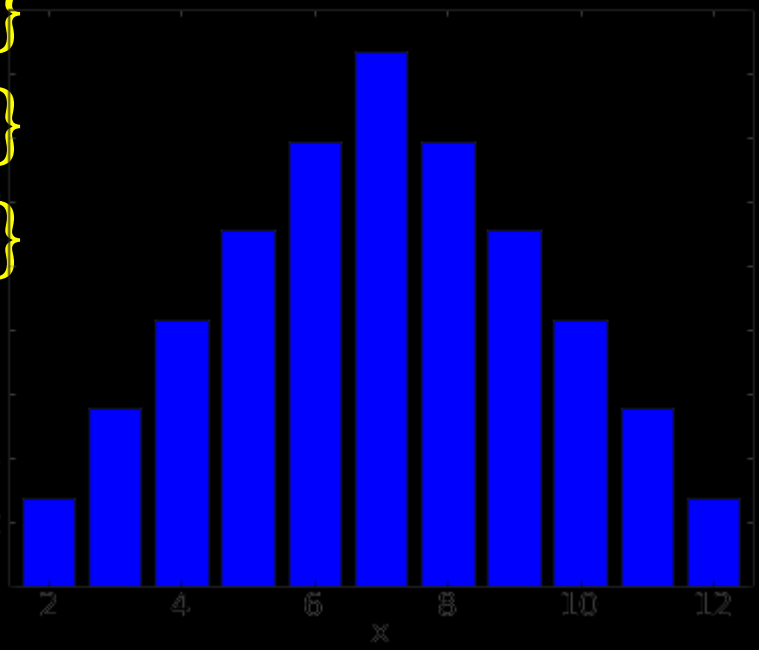
Ex: The probability density function of rolling of a dice is:

$$P(x) = \begin{cases} \frac{1}{6} & if\ x \in \{1,2,3,4,5,6\} \\ 0 & otherwise \end{cases}$$

Ex: Consider X to be the sum on the upper face when two dice are rolled.

Sample space is: {2,3,4,5,6,7,8,9,10,11,12}

$$P(x) = \begin{cases} {}^{1}/_{36} \ if \ x \in \{2,12\} \\ {}^{2}/_{36} \ if \ x \in \{3,11\} \\ {}^{3}/_{36} \ if \ x \in \{4,10\} \\ {}^{4}/_{36} \ if \ x \in \{5,9\} \\ {}^{5}/_{36} \ if \ x \in \{6,8\} \\ {}^{6}/_{36} \quad if \ x = 7 \\ 0 \quad otherwise \end{cases}$$



Continuous probability distribution

Continuous random variable is a random variable that can take on any value from a continuum, e.g. set of all real numbers or an interval. Instead of sum, we take the integral.

The probability distribution function in this case is called a **probability density function**, which assigns the probability that X is near each value.

Given the probability density function ρ(x) for X, we determine the probability that X is in any set A (i.e., that X∈A) by integrating ρ(x) over the set A, i.e.,

$$\Pr(X \in A) = \int_A \rho(x)dx$$

Note that, for continuous random variables, If the set A contains just a single element, then probability that X is equal to that one value is exactly zero, as the integral over a single point is zero.

Also, $\rho(x) \geq 0 \quad for\ all\ x$

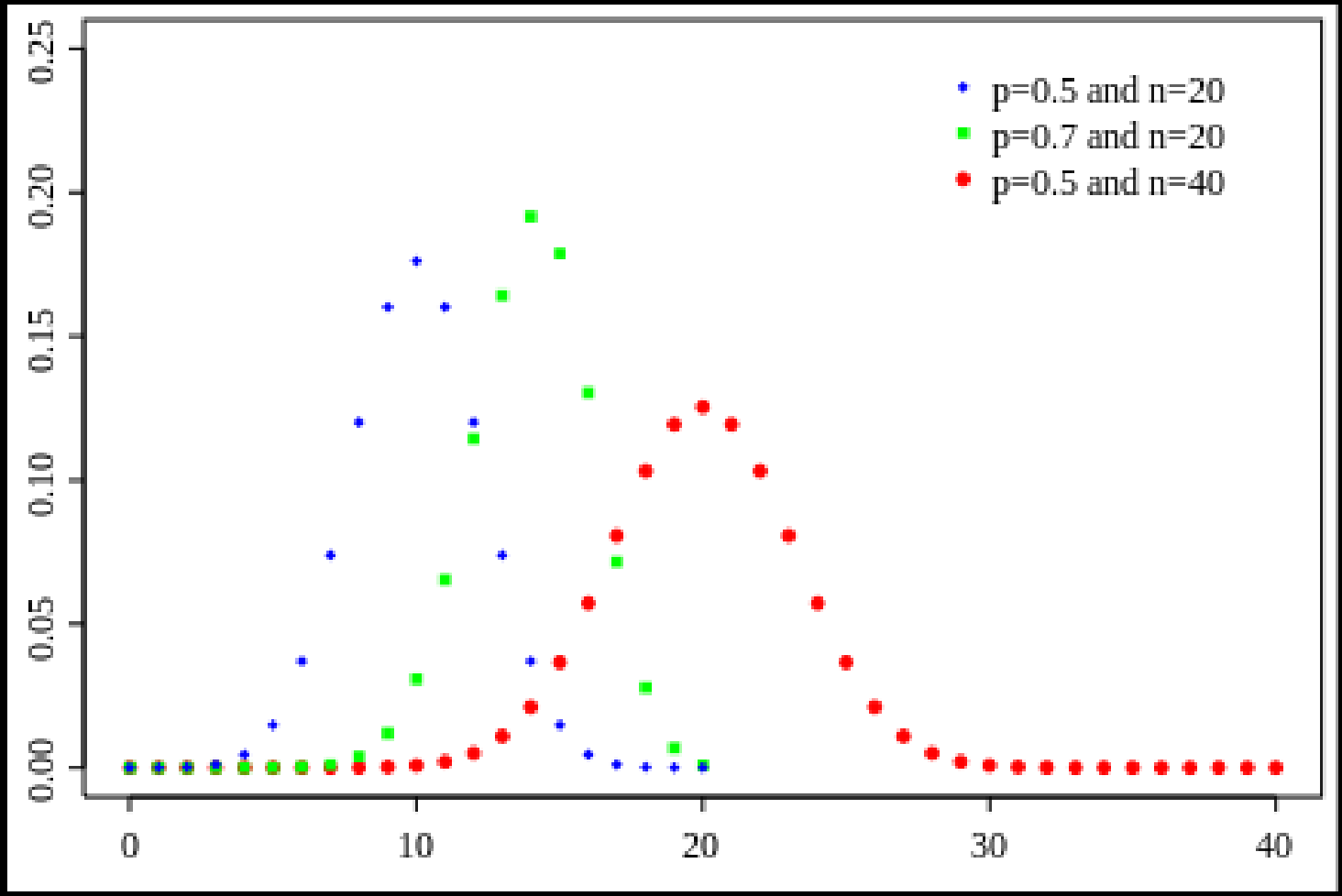
And $\int_A \rho(x)dx = 1$

Binomial Distribution

Def: A **binomial distribution** (bi- meaning two) can be thought of as the probability of a SUCCESS or FAILURE outcome in an experiment which is repeated multiple times. The binomial is a type of distribution that has two possible outcomes.

Ex 1: Tossing of a coin has only two possible outcomes: heads or tails.

Ex 2: Appearing in a driving test has two possible outcomes: pass or fail.


p=0.5 and n=20
p=0.7 and n=20
p=0.5 and n=40
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40

n is the number of times the experiment is conducted

p is the probability of one specific outcome

Ex: To find the probability of getting a 5 on a 6-faced dice.

If the dice is thrown 30 times, the probability of getting 5 on any throw is 1/6. Throw the dice 30 times. This follows a binomial distribution of (n=30, p=1/6) where SUCCESS means out come is 5 and FAILURE means outcome is 1,2,3,4 or 6.

Ex 2: If SUCCESS is defined as getting an even number, then the probability would be (n=20, p=1/2).

Conditions:

1. The experiment has exactly **two** outcomes.
2. **The number of observations or trials is fixed.**
3. **Each observation or trial is independent.**
   That is, none of your trials have an effect on the probability of the next trial.
4. The **probability of success** (tails, heads, fail or pass) is **exactly the same** from one trial to another.

How to identify if an experiment is a Binomial

Which of the following are binomial experiments?

1. An email survey to find out a group of 500 people if they were affected by nCOVID-19 ?
2. Counting the average number of cars arriving at a service centre.
3. Survey 25 traffic lights at 9 am in Delhi to find if the lights are red, orange or green.
4. In a Diwali mela, you play a game to “knock a glass” with 5 balls. There are 20 glasses, out of which 10 have a chit saying “win” and 10 have a chit saying “Loose”.

Bernoulli Distribution

The Bernoulli distribution is the Binomial distribution with n=1.

A **Bernoulli distribution** is a set of Bernoulli trials.

Each Bernoulli trial has one possible outcome, which is either success (S), or failure(F).

In each trial, the probability of success, P(S) = p, is the same. The probability of failure P(F) = 1 – p.

Also, all Bernoulli trials are independent from each other and the probability of success doesn't change from trial to trial, even if information about the other trials' outcomes is known.

Ex: A new vaccine is found to cure nCOVID-19.

Either it cures the disease (successful) or it does not cure the disease (failure).

Ex 2: Buy a lottery ticket. Either you win or you loose.

Ex 3: Appear for an exam, either you pass or you fail.

Formula for binomial distribution

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where:

b = binomial probability

x = total number of "successes"

p = probability of a success on a single trial

n = number of trials

Ex 1: A coin is tossed 15 times. What is the probability of getting exactly 5 heads?

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Sol:

x = 5, p = 0.5, n = 15.

$$\mathrm{P}(x = 5) = \binom{15}{5} 0.5^5 (1-0.5)^{15-5}$$

$$= 3003 \times 0.03125 \times 0.0009765625$$

$$= 0.09164405$$

Ex 2: 80% of people who like to eat out are IITR students. If 9 people who like to eat out are randomly selected, find the probability that exactly 6 are IITR students.

Sol:

$p =$ probability of success = 80 % = 0.8

$1 - p =$ probability of failure = 20% = 0.2

$n = 9$ and $x = 6$

Then, required probability is = $\binom{9}{6} 0.8^6 (1 - 0.8)^{9-6}$

$= 84 \times 0.262144 \times 0.008 = 0.176$

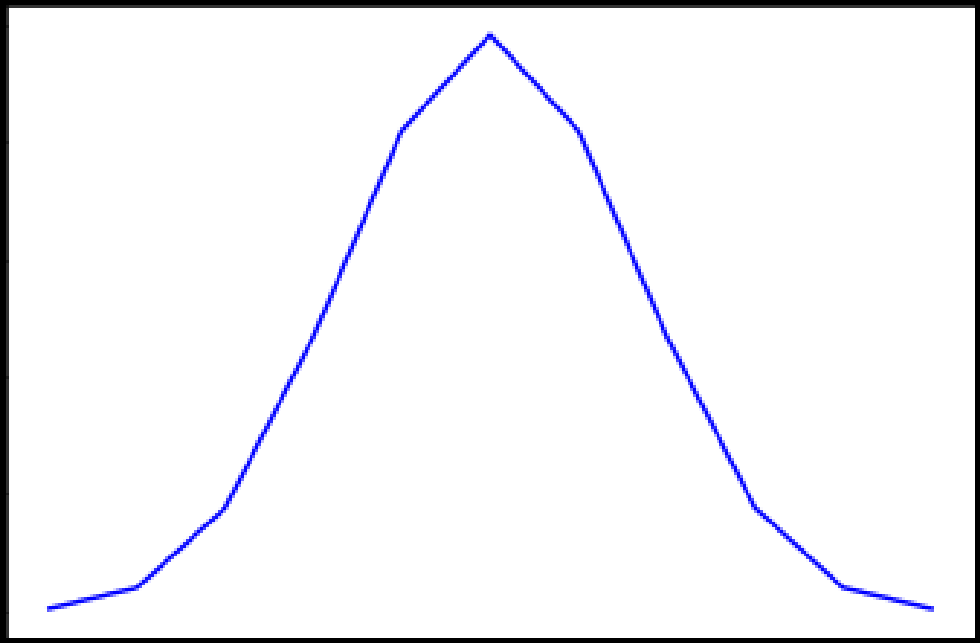
Python code

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom
n = 10
p = 0.5
x = np.arange(0, n+1)
binomial_pmf = binom.pmf(x, n, p)
print(binomial_pmf)
```

Values obtained

[0.00097656 0.00976563 0.04394531 0.1171875
0.20507812 0.24609375 0.20507812 0.1171875
0.04394531 0.00976563 0.00097656]

```
plt.plot(x, binomial_pmf, color='blue')
plt.title(f"Binomial Distribution (n={n}, p={p})")
plt.xlabel("no. of trials")
plt.ylabel("probability")
plt.show()
```



$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$E[x] = np$$

$$var[x] = np(1-p)$$

$$mode[x] = \lfloor (n+1)p \rfloor$$

Where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

$\lfloor (n+1)p \rfloor$ is largest integer $\leq (n+1)p$

Poison Distribution

Def: It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

Ex 1: A call center receives an average of 180 calls per hour, 24 hours a day. The calls are independent; receiving one does not change the probability of when the next one will arrive. The number of calls received during any minute has a Poisson probability distribution with mean 3.

Poison Distribution

It predicts the probability of the occurrence of an event in a fixed time interval.

Ex 2: A travel agency hires 200 taxis every weekend. How to predict how many taxis will be hired in the coming weekend.

Ex 3: If the average number of customers arriving in a restaurant is 500 per week, then will there be more customers in the coming week?

Advantages in business:

1. Estimate the time when the demand is high.
2. Too less stocks means loss in business opportunity.
3. Overstocking may mean losses for unsold goods.
4. Be well prepared for more number of arriving customers, by hiring more staff, more supplies,
5. Prevent wastage of resources.
6. Adjust supply and demand.

Formula for Poison distribution

$$P(x, \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Ex: The average number of landslides that occur in Uttarakhand is 2 per monsoon. What is the probability that exactly 3 landslides will hit Uttarakhand next year?

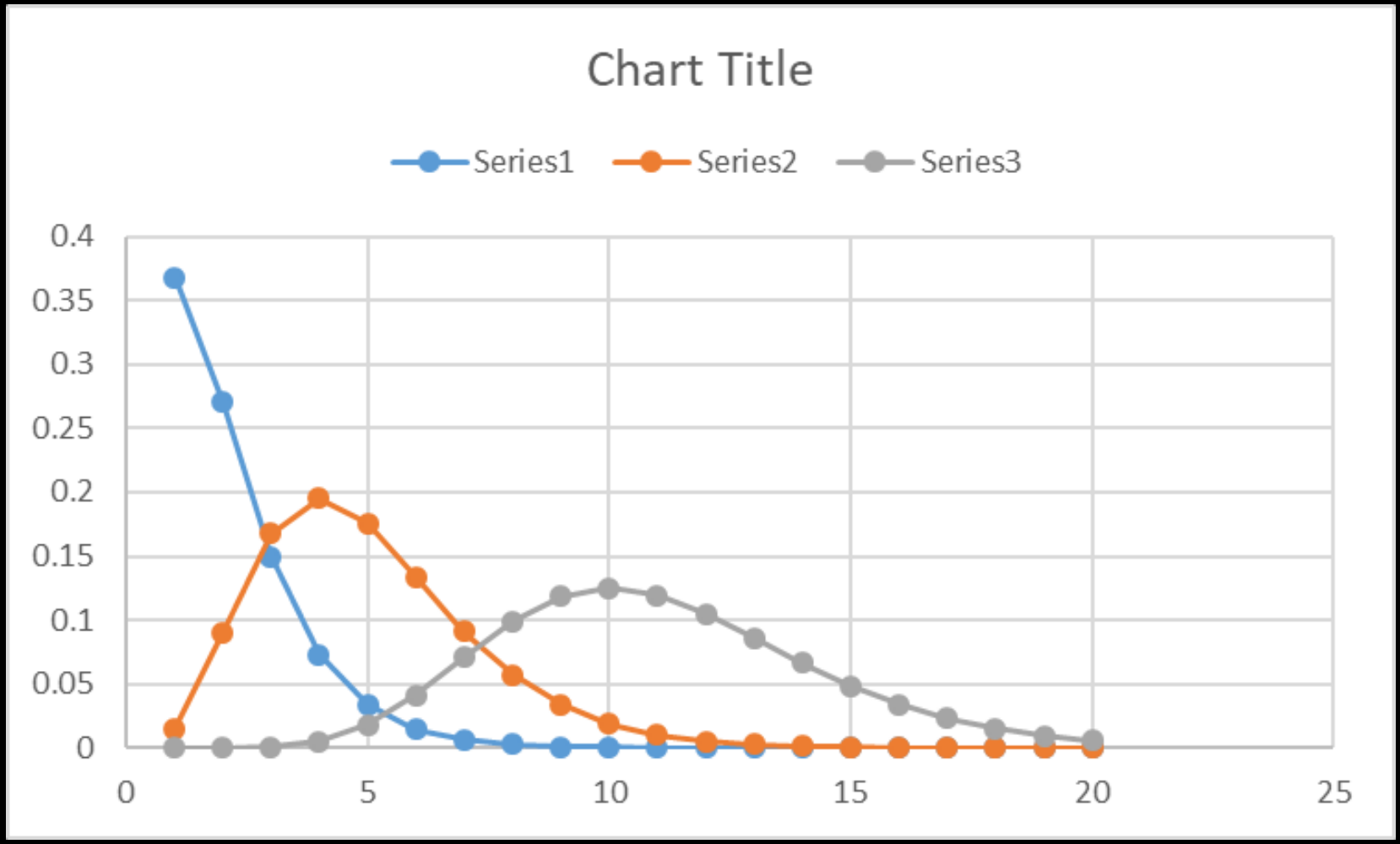
Sol:

$\lambda = 2\ , x = 3\ and\ e = 2.71828$

So, required probability is = $\frac{e^{-2}2^{3}}{3!} = 0.1804$

That is 18 %

Poison distribution for $\lambda = 1,4,10$



Python code

Ex: A store sells 3 apples per day on average. What is the probability that they will sell 5 apples on a given day?

```
from scipy.stats import poisson

#calculate probability

poisson.pmf(k=5, mu=3)
```

Output is:

0.100819

Poison vs Binomial

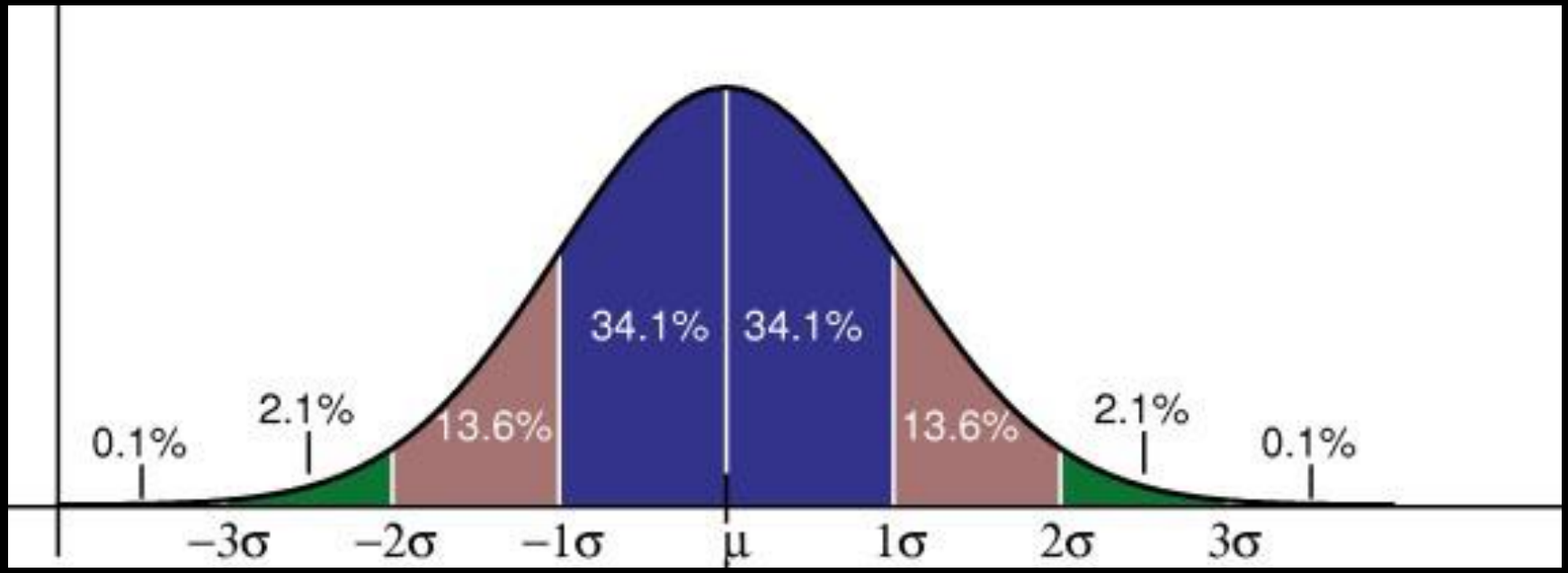
- If situation has an average probability of the occurrence of an event and it is required to find probability of the occurrence of a certain number of events, then use Poison Distribution.
- If the situation gives the exact probability and it is required to find the probability of occurrence of an event certain number of times out of x , then use Binomial Distribution.

Normal Distribution

Def: Normal Distribution / bell curve / De Moivre distribution, is a distribution that occurs naturally in many situations.

Majority will score average (C), while smaller numbers of students will score a B or D, while further smaller percentage of students will score an F or an A. It is symmetrical curve about the mean.

Exs: Results of Tests like JEE, GATE, SAT and GRE.
Weight of students of a class.
Measurement errors.


34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ

Properties of Normal Distribution

- Mean, mode and median are all equal.
- Curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

Interpretation

Ex: If Ram gets 90 in Maths and 2 above standard deviation whereas he gets 95 in English which is one standard deviation above mean, in which subject has he performed better?

Based on the above data, Ram performed better in Math than in English.

Ex: Find the number of houses priced between $50K and 200K out of a sample of 400 houses.

Step 1: Find mean (average or μ) and Standard deviation (σ).

Step 2: Draw a graph and put the mean in the center.

Name of the probability distribution	Probability distribution function	Mean	Variance
Binomial distribution	$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Geometric distribution	$\Pr(X = k) = (1-p)^{k-1} p$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$
Normal distribution	$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Uniform distribution (continuous)	$f(x \mid a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential distribution	$f(x \mid \lambda) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Poisson distribution	$f(k \mid \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$	λ	λ

Probability Density Function Examples

Question 1

Suppose that X is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the value of C?

(b) Find $P\{X > 1\}$.

Solution (a) Since f is a probability density function, we must have $\int_{-\infty}^{\infty} f(x)\,dx = 1$, implying that

$$C\int_0^2 (4x - 2x^2)\,dx = 1$$

or

$$C\left[2x^2 - \frac{2x^3}{3}\right]\Bigg|_{x=0}^{x=2} = 1$$

or

$$C = \frac{3}{8}$$

Hence,

(b) $P\{X > 1\} = \int_1^{\infty} f(x)\,dx = \frac{3}{8}\int_1^2 (4x - 2x^2)\,dx = \frac{1}{2}$ ■

Question 2

The amount of time in hours that a computer functions before breaking down is a continuous random variable with probability density function given by

$$f(x) = \begin{cases} \lambda e^{-x/100} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

What is the probability that

(a) a computer will function between 50 and 150 hours before breaking down?

(b) it will function for fewer than 100 hours?

Solution (a) Since

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \lambda \int_{0}^{\infty} e^{-x/100}\,dx$$

we obtain

$$1 = -\lambda(100)e^{-x/100}\Big|_0^{\infty} = 100\lambda \quad \text{or} \quad \lambda = \frac{1}{100}$$

Hence, the probability that a computer will function between 50 and 150 hours before breaking down is given by

$$\begin{aligned} P\{50 < X < 150\} &= \int_{50}^{150} \frac{1}{100} e^{-x/100}\,dx = -e^{-x/100}\Big|_{50}^{150} \\ &= e^{-1/2} - e^{-3/2} \approx .383 \end{aligned}$$

(b) Similarly,

$$P\{X < 100\} = \int_0^{100} \frac{1}{100} e^{-x/100}\, dx = -e^{-x/100}\Big|_0^{100} = 1 - e^{-1} \approx .632$$

In other words, approximately 63.2 percent of the time, a computer will fail before registering 100 hours of use. ■

Question 3

Find $E[X]$ when the density function of X is

$$f(x) = \begin{cases} 2x & \text{if} \quad 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution

$$
\begin{aligned}
E[X] &= \int xf(x)\,dx \\
&= \int_0^1 2x^2\,dx \\
&= \frac{2}{3}
\end{aligned}
$$

Question 4

For some constant c, the random variable X has the probability density function

$$f(x) = \begin{cases} cx^4 & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find (a) $E[X]$ and (b) $\text{Var}(X)$.

First, let us find c by using

$$1 = \int_0^2 cx^4 dx = 32c/5 \Rightarrow c = 5/32$$

(a) $E[X] = \frac{5}{32}\int_0^2 x^5 dx = \frac{5}{32}\frac{64}{6} = 5/3$

(b) $E[X^2] = \frac{5}{32}\int_0^2 x^6 dx = \frac{5}{32}\frac{128}{7} = 20/7 \Rightarrow \text{Var}(X) = 20/7 - (5/3)^2 = 5/63$

Question 5

5.4. The random variable X has the probability density function

$$f(x) = \begin{cases} ax + bx^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

E(X)=3/5, find a and b

Since

$$1 = \int_0^1 (ax + bx^2)dx = a/2 + b/3$$

$$.6 = \int_0^1 (ax^2 + bx^3)dx = a/3 + b/4$$

we obtain $a = 3.6, b = -2.4$. Hence,

(a) $P\{X < 1/2\} = \int_0^{1/2}(3.6x - 2.4x^2)dx = (1.8x^2 - .8x^3)\Big|_0^{1/2} = .35$

(b) $E[X^2] = \int_0^1 (3.6x^3 - 2.4x^4)dx = .42 \Rightarrow \text{Var}(X) = .06$

Probability Distribution Function - Examples

Ex 1: Uniform Distribution

5.7. To be a winner in a certain game, you must be successful in three successive rounds. The game depends on the value of U, a uniform random variable on $(0, 1)$. If $U > .1$, then you are successful in round 1; if $U > .2$, then you are successful in round 2; and if $U > .3$, then you are successful in round 3.

(a) Find the probability that you are successful in round 1.

(b) Find the conditional probability that you are successful in round 2 given that you were successful in round 1.

(c) Find the conditional probability that you are successful in round 3 given that you were successful in rounds 1 and 2.

(d) Find the probability that you are a winner.

Sol:

$$f(u) = \begin{cases} \frac{1}{1-0}, 0 < u < 1 \\ 0, \quad otherwise \end{cases}$$

$$(a) P\{U > 0.1\} = \int_{0.1}^{1} f(u)du = \int_{0.1}^{1} du = u|_{0.1}^{1} = 1 - 0.1 = 0.9$$

$$(b) P\{U > 0.2|U > 0.1\} = \frac{P\{U>0.2\} \cap P\{U>0.1\}}{P\{U>0.1\}} = \frac{P\{U>0.2\}}{P\{U>0.1\}}$$

$$P\{U > 0.2\} = \int_{0.2}^{1} f(u)du = u\Big|_{0.2}^{1} = 1 - 0.2 = 0.8$$

$$\Rightarrow P\{U > 0.2|U > 0.1\} = \frac{0.8}{0.9} = \frac{8}{9}$$

Sol:

(c) $P\{U > 0.3 | U > 0.2, U > 0.1\} = \frac{P(U>0.3, U>0.2, U>0.1)}{P(U>0.2, U>0.1)}$

$$\Rightarrow \frac{P(U > 0.3)}{P(U > 0.2)}$$

$$P(U > 0.3) = \int_{0.3}^{1} f(u)du = 1 - 0.3 = 0.7$$

$$P\{U > 0.3 | U > 0.2, U > 0.1\} = \frac{0.7}{0.8} = \frac{7}{8}$$

(d) $P\{U > 0.3\} = P(U > 0.1) \times P\{U > 0.2 | U > 0.1\} \times P\{U > 0.3 | U > 0.2, U > 0.1\} = \frac{9}{10} \times \frac{8}{9} \times \frac{7}{8} = \frac{7}{10}$

Ex 2: Normal Distribution

Suppose that X is a normal random variable with mean 5. If $P\{X > 9\} = .2$, approximately what is $\mathrm{Var}(X)$?

Sol:

mean = 5

$$P\left(\frac{x-5}{\sigma} > \frac{9-5}{\sigma}\right) = 0.2$$

$$\Rightarrow 1 - \Phi\left(\frac{4}{\sigma}\right) = 0.2$$

$$\Rightarrow \Phi\left(\frac{4}{\sigma}\right) = 1 - 0.2 = 0.8$$

$$\Rightarrow \frac{4}{\sigma} \approx 0.84$$

$$\Rightarrow \sigma \approx \frac{4}{0.84} \approx 4.76$$

$$\sigma^2 \approx 22.66$$

Ex 3: Normal Distribution

Let X be a normal random variable with mean 12 and variance 4. Find the value of c such that $P\{X > c\} = .10$.

Sol:

mean = 12, var = 4, std = 2

$$P\left(\frac{X-12}{2} > \frac{c-12}{2}\right) = 0.10$$

$$\Rightarrow 1 - \Phi\left(\frac{c-12}{2}\right) = 0.10$$

$$\Rightarrow \frac{c-12}{2} = \Phi^{\{-1\}}(0.90) \approx 1.28$$

$$\Rightarrow c \approx 2 \times 1.28 + 12 \approx 14.56$$

Ex 4: Normal Distribution

If X is a normal random variable with parameters $\mu = 10$ and $\sigma^2 = 36$, compute

(a) $P\{X > 5\}$;

(b) $P\{4 < X < 16\}$;

(c) $P\{X < 8\}$;

(d) $P\{X < 20\}$;

(e) $P\{X > 16\}$.

Sol:

$$(a) P\{X > 5\} = P\left\{\frac{X-10}{6} > \frac{5-10}{6}\right\} = P\{Z \succ -0.833\}$$

$$= 1 - \Phi(-0.833) = 1 - (1 - \Phi(0.833))$$

$$\approx 0.9616$$

$$\text{(b)} P\{4 < X < 16\} = P\left\{\frac{4-10}{6} < X < \frac{16-10}{6}\right\} = \Phi(1) - \Phi(-1)$$

$$\Phi(1) - \big(1 - \Phi(1)\big) = 2.\Phi(1) - 1$$

$$\approx 2(0.8413) - 1 = 0.6826$$

$$\text{(c) } P\{X < 8\} = P\left\{X < \frac{8-10}{6}\right\} = \Phi\left(-\frac{1}{3}\right) = 1 - \Phi\left(\frac{1}{3}\right) \approx 1 - 0.6293$$

$$= 0.37070$$

(d) And (e) are assignment

Ex 5: Exponential Distribution

On the average a certain computer part lasts ten years. The length of time the computer part lasts is exponentially distributed.

- What is the probability that a computer part lasts more than 7 years?

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases}$$

Sol:

$$\mu = 10$$

$$\lambda = \frac{1}{\mu} = 0.1$$

$F(a) = P\{X \leq a\} = \int_0^a \lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda a}$, for $a \geq 0$

Now, $P\{X > 7\} = 1 - P\{X \leq 7\} = 1 - F(a)$

For exponential distribution function $F(a) = 1 - e^{-\lambda a}$.

Thus, $P\{X > 7\} = 1 - \left(1 - e^{-\lambda 7}\right) = e^{-0.1*7} = e^{-0.7} = 0.4966$

Ex 6: Exponential Distribution

- Prove that If X be an exponential random variable with parameter λ.

  Then, $E[X^n] = \frac{n}{\lambda} E[X^{n-1}]$

Sol:

The density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$E[X^n] = \int_0^\infty x^n \lambda\, e^{-\lambda x} dx$$

Integration by substitution $\lambda e^{-\lambda x} = v$ and $u = x^n$

$$E[X^n] = -x^n e^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} n x^{n-1} dx = 0 + \frac{n}{\lambda}\int_0^\infty e^{-\lambda x} x^{n-1} dx$$

$$= \frac{n}{\lambda} E[X^{n-1}]$$

Ex 7: Exponential Distribution

5.32. The time (in hours) required to repair a machine is an exponentially distributed random variable with parameter $\lambda = \frac{1}{2}$. What is

(a) the probability that a repair time exceeds 2 hours?

(b) the conditional probability that a repair takes at least 10 hours, given that its duration exceeds 9 hours?

Sol:

$$\lambda = \frac{1}{2}$$

$$(a) P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - \left(1 - e^{-2\lambda}\right) = e^{-1}.$$

$$\text{(b) } P(X \geq 10 | X > 9) = \frac{P(X \geq 10, X > 9)}{P(X > 9)} = \frac{P(X \geq 10)}{P(X > 9)} = \frac{e^{-10}}{e^{-9}} = e^{-1}$$

Ex 8: Geometric Distribution

A coin is biased so that the probability of obtaining “heads” in any toss is $p, p \neq \frac{1}{2}$. The coin is tossed repeatedly until a “head" is obtained. The sum of probability of obtaining “heads” after an even number of tosses is $\frac{2}{5}$. Determine the value of p .

P{X=n} = (1-p)^(n-1)p^1

Sol:

Let p be the probability of getting head
Let X be the number of tossed until head come

$P(X = 2) = (1 - p)p$
$P(X = 4) = (1 - p)^3 p$
$P(X = 6) = (1 - p)^5 p$

Now, according to question,

$$P(X = 2) + P(X = 4) + P(X = 6) + \cdots = \frac{2}{5}$$

$$(1 - p)p + (1 - p)^3 p + (1 - p)^5 p + \cdots = \frac{2}{5}$$

Sol:

$$(1-p)p + (1-p)^3p + (1-p)^5p + \cdots = \frac{2}{5}$$

Series is in G.P. with a = $(1-p)p$ and r = $(1-p)^2$.

Now of sum of G.P. of infinity series is

$\sum_n ar^n = \frac{a}{1-r}$.

$$\frac{(1-p)p}{1-(1-p)^2} = \frac{(1-p)p}{2p-p^2} = \frac{1-p}{2-p} = \frac{2}{5}$$

On solving , $p = \frac{1}{3}$

Hypothesis Testing

It is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

Its purpose is to determine whether there is enough statistical evidence in favor of a certain belief, or hypothesis, about a parameter.

There are 5 steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis and alternate hypothesis H_0 and H_a or H_1.
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.

Step 1: State your null and alternate hypothesis

After developing your initial research hypothesis (the prediction that you want to investigate), it is important to restate it as a null (Ho) and alternate (Ha) hypothesis so that you can test it mathematically.

The alternate hypothesis is usually your initial hypothesis that predicts a relationship between variables. The null hypothesis is a prediction of no relationship between the variables you are interested in.

You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:

Ho: Men are, on average, not taller than women.

Ha: Men are, on average, taller than women.

Step 2: Collect data

For a statistical test to be valid, it is important to perform sampling and collect data in a way that is designed to test your hypothesis. If your data are not representative, then you cannot make statistical inferences about the population.

To test differences in average height between men and women, your sample should have an equal proportion of men and women, and cover a variety of socio-economic classes and any other control variables that might affect average height.

You should also consider your scope (Worldwide / country / city). A potential data source in this case might be census data, since it includes data from a variety of regions and social classes and is usually available.

Step 3: Perform a statistical test based on the comparison of :

(1). within-group variance (2). Between-group variance.

If the between-group variance is large enough that there is little or no overlap between groups, then your statistical test will reflect that by showing a l**ow p-value.** This means it is unlikely that the differences between these groups came about by chance.

Alternatively, if there is high within-group variance and low between-group variance, then your statistical test will reflect that with a **high p-value.** This means it is likely that any difference you measure between groups is due to chance.

Your choice of statistical test will be based on the type of data you collected.

Based on the type of data you collected, you perform a one-tailed t-test to test whether men are in fact taller than women. This test gives you:

- an estimate of the difference in average height between the two groups.
- a p-value showing how likely you are to see this difference if the null hypothesis of no difference is true.

Your t-test shows an average height of 175.4 cm for men and an average height of 161.7 cm for women, with an estimate of the true difference ranging from 10.2cm to infinity. The p-value is 0.002.

Step 4: Decide whether to **reject** or **fail to reject** your null hypothesis based on the outcome of your statistical test.

In most cases you will use the p-value generated by your statistical test to guide your decision. And in most cases, your predetermined level of significance for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis (Type I error).

In your analysis of the difference in average height between men and women, you find that the p-value of 0.002 is below your cutoff of 0.05, so you decide to reject your null hypothesis of no difference.

Step 5: Present your findings

Give a brief summary of the data and a summary of the results of your statistical test (for example, the estimated difference between group means and associated p-value). You can discuss whether your initial hypothesis was supported by your results or not.

In the formal language of hypothesis testing, we talk about rejecting or failing to reject the null hypothesis.

What is p-value

Def: A **p-value**, or probability value, is a number describing how likely it is that your data would have occurred under the null hypothesis of your statistical test.

How to calculate it

Automatically calculated by software or can be estimated using p-value tables for the relevant test statistic.

It is calculated from the null distribution of the test statistic. They tell you how often a test statistic is expected to occur under the null hypothesis of the statistical test, based on where it falls in the null distribution.

If the test statistic is far from the mean of the null distribution, then the p-value will be small, showing that the test statistic is not likely to have occurred under the null hypothesis.

P-value	Decision
P-value > 0.05	The result is not statistically significant and hence don’t reject the null hypothesis.
P-value < 0.05	The result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis.
P-value < 0.01	The result is highly statistically significant, and thus rejects the null hypothesis in favour of the alternative hypothesis.

Ex: Flip a coin ten times with the null hypothesis that it is fair. The total number of heads is the test statistic, which is two-tailed. Assume that alternating heads and tails on each flip is observed. (HTHTHTHTHT). As this is the predicted number of heads, the test statistic is 5 and the p-value is 1 (totally unexceptional).

Assume that the test statistic was the “number of alternations” (i.e., the number of times H followed T or T followed H), which is also two-tailed. This would result in a test statistic of 9, which is extremely high and has a p-value of $^1/_{2^8} = {}^1/_{256}$ or roughly 0.0039. This would be regarded as extremely significant, much beyond 0.05 level. These findings suggest that the data set is exceedingly improbable to have happened by random in terms of one test statistic, yet they do not imply that the coin is biased towards heads or tails.

According to the first statistic, the data have a high p-value, indicating that the number of heads observed is not impossible.

According to the second test statistic, the data have a low p-value indicating that the pattern of flips observed is extremely unlikely.

There is no “alternative hypothesis,” (therefore only the null hypothesis can be rejected).

Reasons could be – the data could be falsified, or the coin could have been flipped by a magician who purposefully swapped outcomes.

Thus the p-value is entirely dependent on the test statistic used and that p-values can only be used to reject a null hypothesis, not to explore an alternate hypothesis.

P-value Formula

P-value lies between 0 and 1.Level of significance(α) is a predefined threshold usually 0.05.

Step 1: Find out the test static $z = \frac{\hat{p} - p0}{\sqrt{\frac{p0(1-p0)}{N}}}$

Where:

$\hat{p}$ is Sample Proportion

$p0$ is assumed population proportion in the null hypothesis

N is sample size

Step 2: Look at the Z-table to find corresponding level of P from the z value obtained.

Ex: A statistician wants to test the hypothesis H_0: μ = 120 using the alternative hypothesis Hα: μ > 120 and assuming that α = 0.05, using sample values as: n =40, σ = 32.17 and x̄ = 105.37. Determine the conclusion for this hypothesis?

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{32.17}{\sqrt{40}} = 5.0865$$

Using test static formula,

$$t = \frac{105.37 - 120}{5.0865} = -2.8762$$

Using Z-Score table, find the value of $P(t > -2.8762)$.

From the table, we get

$$P(t < -2.8762) = P(t > -2.8762) = 0.003$$

So, If $P(t > -2.8762) = 1 - 0.003 = 0.997$

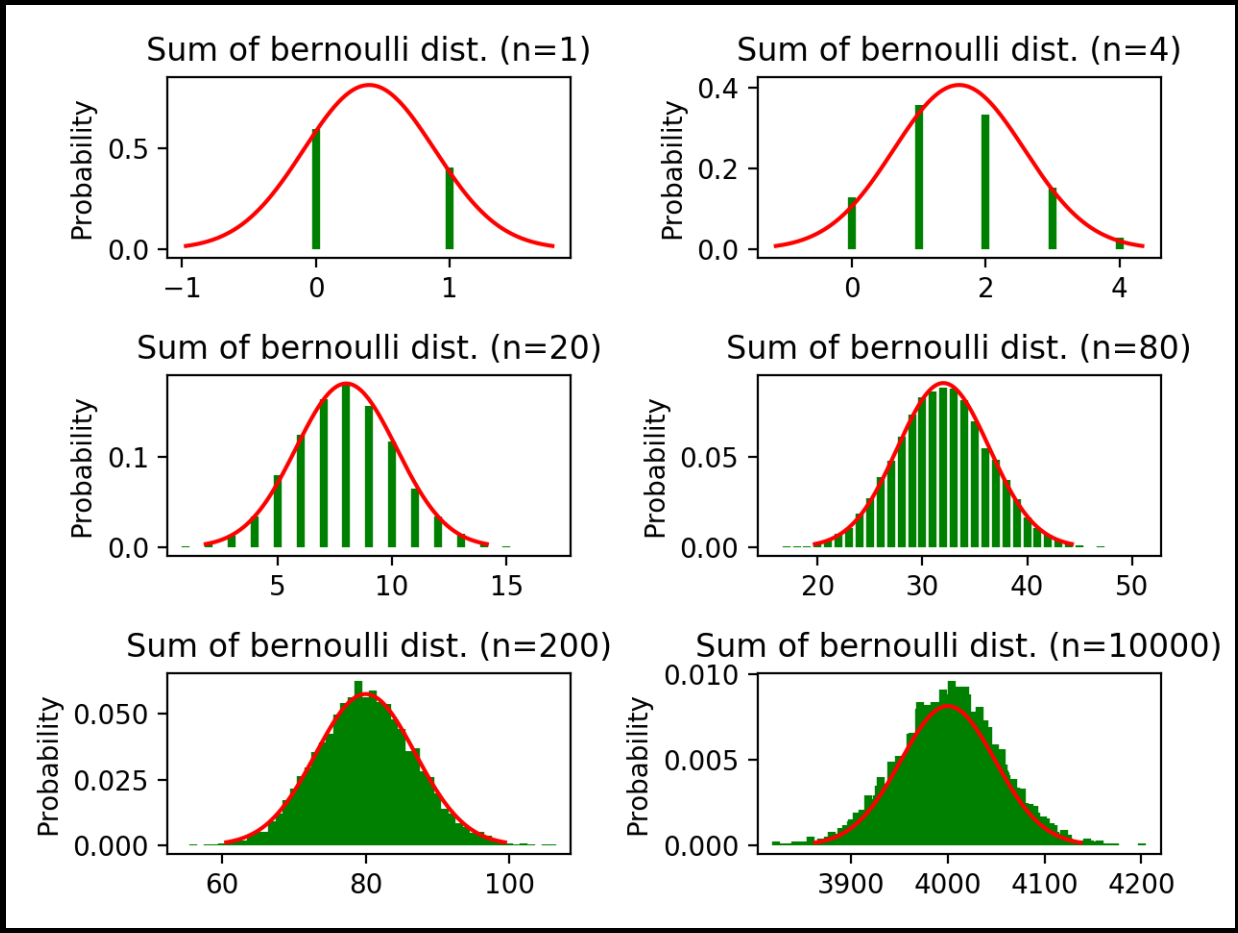
P-value = 0.997 > 0.05

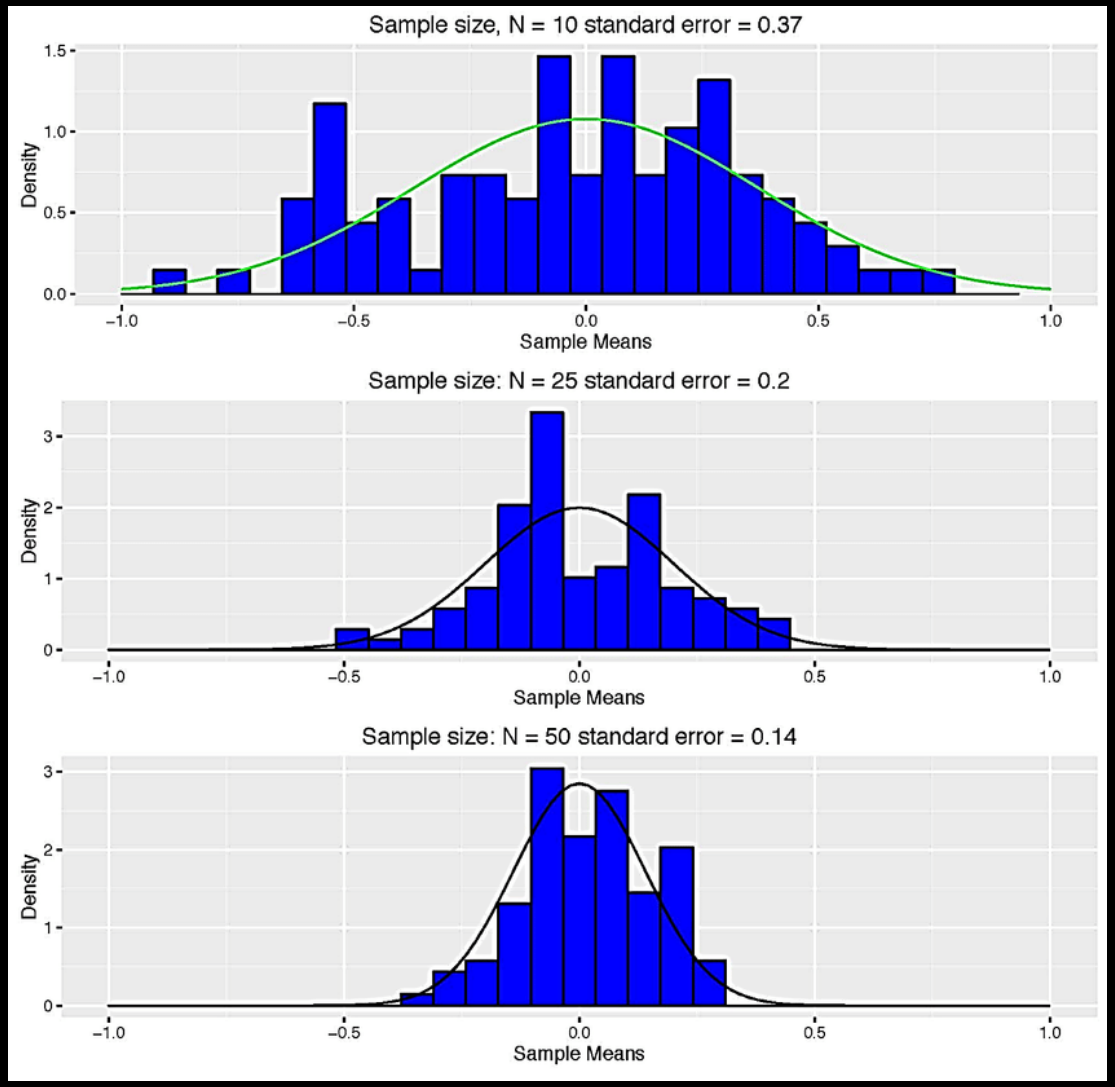
Therefore, from the conclusion, if p>0.05, the null hypothesis is accepted or fails to reject.

Hence, the conclusion is "fails to reject H_0."

Central Limit Theorem

The central limit theorem states that if there is a population with mean μ and standard deviation σ and sufficiently large random samples are taken from the population with replacement, then the distribution of the sample means will be approximately normally distributed.


Sum of bernoulli dist. (n=1)
Probability
0.5
0.0
−1
0
1
Sum of bernoulli dist. (n=4)
Probability
0.4
0.2
0.0
0
2
4
Sum of bernoulli dist. (n=20)
Probability
0.1
0.0
5
10
15
Sum of bernoulli dist. (n=80)
Probability
0.05
0.00
20
30
40
50
Sum of bernoulli dist. (n=200)
Probability
0.050
0.025
0.000
60
80
100
Sum of bernoulli dist. (n=10000)
Probability
0.010
0.005
0.000
3900
4000
4100
4200


Sample size, N = 10 standard error = 0.37
Density
Sample Means
Sample size: N = 25 standard error = 0.2
Density
Sample Means
Sample size: N = 50 standard error = 0.14
Density
Sample Means

Ex: An investor is interested in estimating the return of ABC stock market index that is comprised of 100,000 stocks. Due to the large size of the index, the investor is unable to analyze each stock independently and instead chooses to use random sampling to get an estimate of the overall return of the index.

The investor picks random samples of the stocks, with each sample comprising at least 30 stocks. The samples must be random, and any previously selected samples must be replaced in subsequent samples to avoid bias.

If the first sample produces an average return of 7.5%, the next sample may produce an average return of 7.8%. With the nature of randomized sampling, each sample will produce a different result. As the size of the sample is increasd with each sample, the sample means will start forming their own distributions.

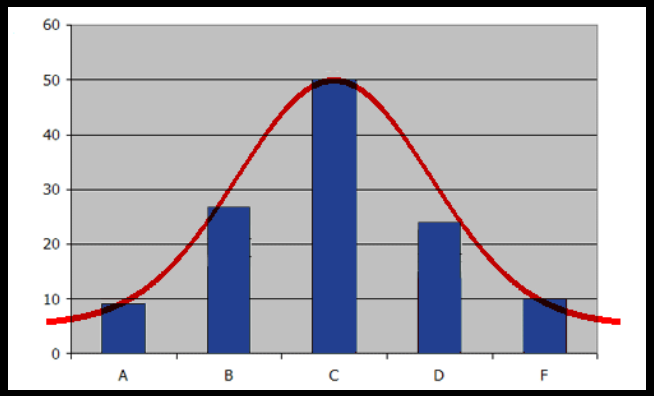
The distribution of the sample means will move toward normal as the value of n increases. The average return of the stocks in the sample index estimates the return of the whole index of 100,000 stocks, and the average return is normally distributed.

Z-test

Def: **Z-test** is a type of hypothesis test, to find out if results from a test are valid or repeatable.

e.g. If someone claims that he has found a new drug that cures Covid. In order to be sure if it is probably true, a hypothesis test is done.

Z test, is used when data is approximately normally distributed



Z test is used if:

- Sample size > 30. Else use a t test.
- Data points should be independent from each other.
- Data is normally distributed. However, for large sample sizes (over 30) this doesn’t always matter.
- Data should be randomly selected from a population, where each item has an equal chance of being selected.
- Sample sizes should be equal if at all possible.

How to run a Z Test?

There are 5 steps:

1. State the null hypothesis and alternate hypothesis.
2. Choose an alpha level.
3. Find the critical value of z in a z table.
4. Calculate the z test statistic (see below).
5. Compare the test statistic to the critical z value and decide if to support or reject the null hypothesis.

You could perform all these steps by hand. For example, you could find a critical value by hand, or calculate a z value by hand. For a step by step example, watch the following video:

Ex: Suppose a company claims that the average time it takes to respond to customer inquiries is 2 minutes. To test this claim, a random sample of 50 customer inquiries is taken, and the average response time is found to be 1.8 minutes with a standard deviation of 0.4 minutes.

We want to test the null hypothesis that the true mean response time is 2 minutes, against the alternative hypothesis that the true mean response time is less than 2 minutes. We can use a one-tailed z-test with a significance level of 0.05 to determine if there is sufficient evidence to reject the null hypothesis.

Sol:

Formula for the z-test statistic is: $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$

where $\bar{x}$ is the sample mean, μ is the hypothesized population mean, σ is the population standard deviation (if known), and n is the sample size.

Substituting the values from the problem, we get:

$z = (1.8 - 2) / (0.4 / \sqrt{50}) = -2.236$

Using a z-table or calculator, we can find that the probability of getting a z-value of -2.236 or lower is approximately 0.012. Since this probability is less than our chosen significance level of 0.05, we can reject the null hypothesis and conclude that there is sufficient evidence to suggest that the true mean response time is less than 2 minutes.

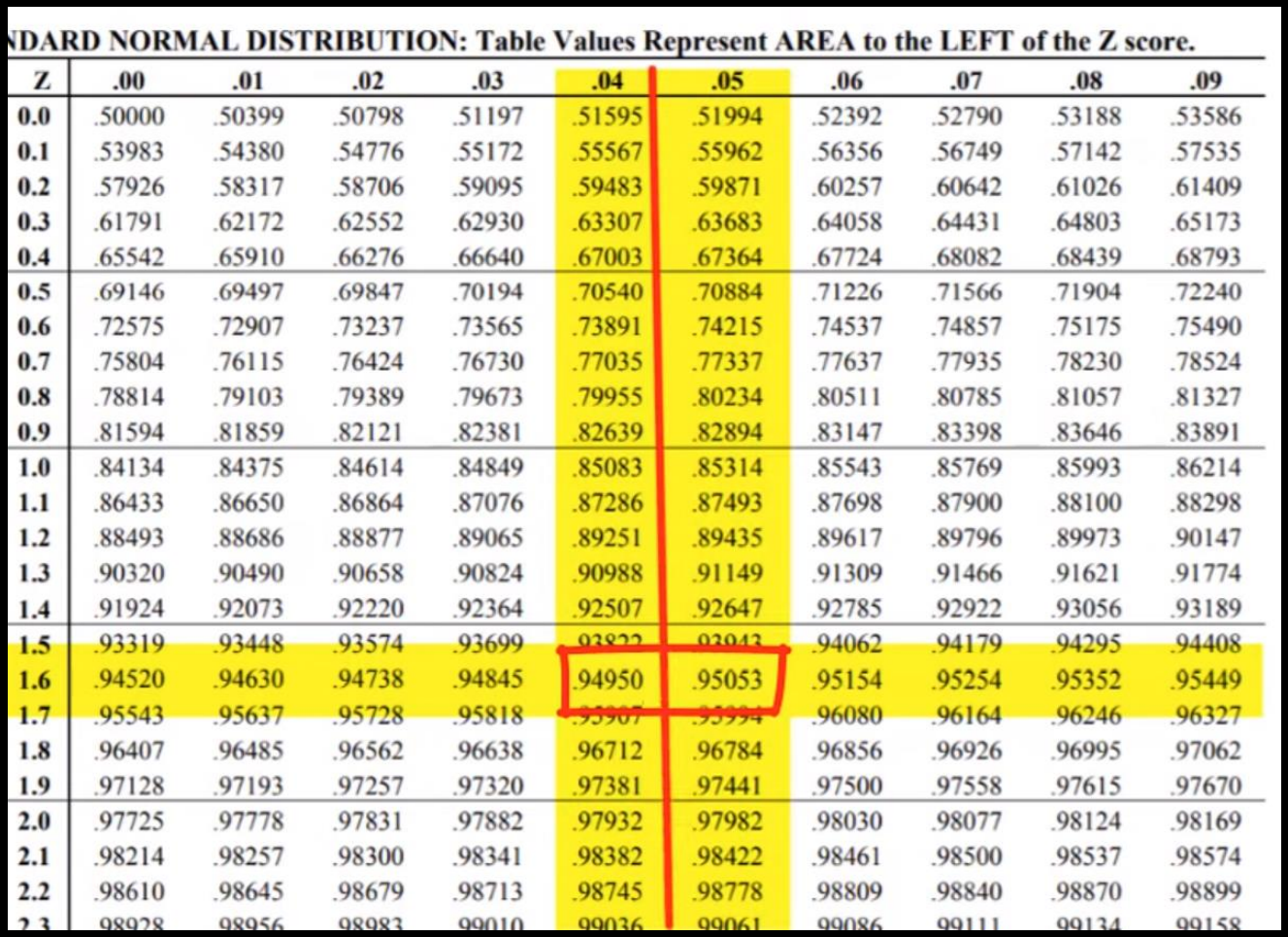
Note that the negative sign of the z-value indicates that the sample mean is less than the hypothesized population mean. Also, since we are using a one-tailed test, we are only interested in the area under the curve to the left of the z-value.

Ex: It is claimed that the students of IITR are above average intelligent. A random sample of 30 IQ scores has a mean score of 112.5. Is there sufficient evidence to support this claim, given that the mean population IQ is 100 with a standard deviation of 15?

Sol:

Null hypothesis H_0: $\mu = 100$ and $H_1: \mu > 100$

$standard\ \alpha - level\ is\ \alpha = 0.05 = 5\%$



NDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	[illegible]	[illegible]	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	[illegible]	[illegible]	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158

1.6+0.045 = 1.645

$$z = \frac{\bar{\bar{x}} - \mu_0}{\sigma\sqrt{n}} = \frac{112.5 - 100}{15\sqrt{30}} = 4.56$$

But 4.56 > 1.645

So we can reject the hypothesis

T-test

Def: t-test is a statistical test that compares the means of two samples. It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

Ex: On flipping a coin 1,000 times, the number of heads follows a normal distribution for all trials. So sample variance fcan be known, but the population variance is unknown.

Ex: Suppose a company claims that the average weight of their product packaging is 500 grams. To test this claim, a random sample of 25 product packages is taken, and the average weight is found to be 485 grams with a standard deviation of 30 grams.

We want to test the null hypothesis that the true mean weight is 500 grams, against the alternative hypothesis that the true mean weight is less than 500 grams. We can use a one-tailed t-test with a significance level of 0.05 to determine if there is sufficient evidence to reject the null hypothesis.

Sol:

Formula for the t-test statistic is: $t = (\bar{x} - \mu) / (s / \sqrt{n})$

where $\bar{x}$ is the sample mean, μ is the hypothesized population mean, s is the sample standard deviation, and n is the sample size.

Substituting the values from the problem, we get:

$t = (485 - 500) / (30 / \sqrt{25}) = -2.5$

Using a t-table or calculator with 24 degrees of freedom (df = n - 1), we can find that the critical t-value for a one-tailed test with a significance level of 0.05 is approximately -1.711.

Since our calculated t-value (-2.5) is less than the critical t-value (-1.711), we can reject the null hypothesis and conclude that there is sufficient evidence to suggest that the true mean weight is less than 500 grams.

Note that the negative sign of the t-value indicates that the sample mean is less than the hypothesized population mean. Also, since we are using a one-tailed test, we are only interested in the area under the curve to the left of the t-value.

Ex 2: A drug company may want to test a new Covid drug to find out if it improves life expectancy. In an experiment, there's always a control group (a group who are given a particular medicine). So while the control group may show an average life expectancy of +5 years, the group taking the new drug might have a life expectancy of +6 years. It would seem that the drug might work. But it could be due to a fluke. To test this, it is advisable to use a Student's t-test to find out if the results are repeatable for an entire population.

In addition, a t test uses a t-statistic and compares this to t-distribution values to determine if the results are statistically significant.

However, use a t test to compare two means, but to compare three or more means, use an ANOVA.

Def: The t score is a ratio between the difference between two groups and the difference within the groups.

Larger t scores = more difference between groups.

Smaller t score = more similarity between groups.

A t score of 3 tells you that the groups are three times as different from each other as they are within each other. So when you run a t test, bigger t-values equal a greater probability that the results are repeatable.

T-Values and P-values

How big is “big enough”? Every t-value has a p-value to go with it. A p-value from a t test is the probability that the results from your sample data occurred by chance. P-values are from 0% to 100% and are usually written as a decimal (for example, a p value of 5% is 0.05). Low p-values indicate your data did not occur by chance. For example, a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance.

There are three main types of t-test:

1. An Independent Samples t-test compares the means for two groups.
2. A Paired sample t-test compares means from the same group at different times (say, one month apart).
3. A One sample t-test tests the mean of a single group against a known mean.

What is an Independent Samples T Test?

The **independent samples t test** (also called the **unpaired samples t test**) helps to compare the means of two sets of data. For example, you could run a t test to see if the average test scores of males and females are different; the test answers the question, “Could these differences have occurred by random chance?”

One sample t test: is used to compare a result to an expected value. For example, do males score higher than the average of 70 on a test if their exam time is switched to 8 a.m.?

Paired t test (**dependent samples**): used to compare related observations. For example, do test scores differ significantly if the test is taken at 8 a.m. or noon?

This test is extremely useful because for the z test you need to know facts about the population, like the population standard deviation, but with the independent samples t test, you don't need to know this information. You should use this test when:

- You do not know the population mean or standard deviation.
- You have two independent, separate samples.

F test

Def: An "**F Test**" uses the F-distribution, so as to compare Two Variances.

1. State the null hypothesis and the alternate hypothesis.
2. Calculate the F value. The F Value is calculated using the formula F = (SSE1 – SSE2 / m) / SSE2 / n-k, where SSE = residual sum of squares, m = number of restrictions and k = number of independent variables.
3. Find the F Statistic (the critical value for this test). The F statistic formula is:

   F Statistic = variance of the group means / mean of the within group variances.

1. You can find the F Statistic in the F-Table.
2. Support or Reject the Null Hypothesis.

F Test to Compare Two Variances

A Statistical F Test uses an F Statistic to compare two variances, s_1 and s_2, by dividing them. The result is always a positive number (because variances are always positive). The equation for comparing two variances with the f-test is:

$$F = \frac{s_1^2}{s_2^2}$$

If the variances are equal, the ratio of the variances will equal 1.

You always test that the population variances are equal when running an F Test. In other words, you always assume that the variances are equal to 1. **Therefore, your null hypothesis will always be that the variances are equal.**

Assumptions:

1. Population must be approximately normally distributed.
2. Samples must be independent events.
3. The larger variance should always go in the numerator to force the test into a right-tailed test, which are easier to calculate.
4. For two-tailed tests, divide alpha by 2 before finding the right critical value.
5. If you are given standard deviations, they must be squared to get the variances.
6. If your degrees of freedom aren't listed in the F Table, use the larger critical value. This helps to avoid the possibility of Type I errors.

Steps to calculate F test

Step 1: If standard deviations is given, go to Step 2. If variances is given, go to Step 3.

Step 2: Square both standard deviations to get the variances.

Step 3: Take the largest variance, and divide it by the smallest variance to get the f-value.

Step 4: Find degrees of freedom , which is sample size minus 1. There will be two degrees of freedom: one for the numerator and one for the denominator.

Step 5: Read out the f-value obtained in Step 3 in the f-table.

Step 6: Compare calculated value (Step 3) with the table f-value in Step 5. If the f-table value is smaller than the calculated value, **reject null hypothesis.**

Two Tailed F-Test

The difference between running a one or two tailed F test is that the alpha level needs to be halved for two tailed F tests.

With a two tailed F test, you just want to know if the variances are not equal to each other.

That is : $H_a = \sigma_1^2 \neq \sigma_2^2$

Ex: Suppose we want to compare the variances of the heights of two different populations. We take a random sample of 25 individuals from each population and calculate the sample variances as $s1^2 = 16$ and $s2^2 = 25$.

We want to test the null hypothesis that the two populations have equal variances, against the alternative hypothesis that they have unequal variances. We can use a two-tailed F-test with a significance level of 0.05 to determine if there is sufficient evidence to reject the null hypothesis.

Sol:

The formula for the F-test statistic is: $F = s1^2 / s2^2$

If the null hypothesis is true, the F-statistic follows an F-distribution with (n1 - 1) and (n2 - 1) degrees of freedom.

Substituting the values from the problem, we get:

$F = 16 / 25 = 0.64$

Using an F-table or calculator with 24 degrees of freedom for both numerator and denominator, we can find that the critical F-value for a two-tailed test with a significance level of 0.05 is approximately 0.36 and 2.78.

Since our calculated F-value (0.64) falls between the critical values of 0.36 and 2.78, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that the variances of the two populations are significantly different.

Note that if the calculated F-value had been greater than the critical value of 2.78 or less than the critical value of 0.36, we would have rejected the null hypothesis in favor of the alternative hypothesis that the two populations have unequal variances. Also, since we are using a two-tailed test, we are interested in the areas under both tails of the F-distribution.

Ex: Conduct a two tailed F Test on following samples:

Sample 1: Variance = 109.63, sample size = 41.

Sample 2: Variance = 65.99, sample size = 21.

Step 1: Write your hypothesis statements:

H_0 = No difference in variances.

H_a = Difference in variances.

Step 2: Calculate your F critical value.

F Statistic = variance 1/ variance 2 = 109.63 / 65.99 = 1.66

Step 3: Calculate the degrees of freedom:

Sample 1 has 40 df (the numerator).

Sample 2 has 20 df (the denominator).

Step 4: No alpha is given so take alpha as 0.05. This needs to be halved for the two-tailed test, so use 0.025.

Step 5: Find the critical F Value using the F Table. Look for alpha = .025 table. Critical F (40,20) at alpha (0.025) = 2.287.

/	df_1=1	2	24	30	40	60	120	∞
df_2=1	647.7890	799.5000	997.2492	1001.414	1005.598	1009.800	1014.020	1018.258
2	38.5063	39.0000	39.4562	39.465	39.473	39.481	39.490	39.498
3	17.4434	16.0441	14.1241	14.081	14.037	13.992	13.947	13.902
4	12.2179	10.6491	8.5109	8.461	8.411	8.360	8.309	8.257
5	10.0070	8.4336	6.3780	6.227	6.175	6.123	6.069	6.015
16	6.1151	4.6867	2.6252	2.568	2.509	2.447	2.383	2.316
17	6.0420	4.6189	2.5598	2.502	2.442	2.380	2.315	2.247
18	5.9781	4.5597	2.5027	2.445	2.384	2.321	2.256	2.187
19	5.9216	4.5075	2.4523	2.394	2.333	2.270	2.203	2.133
20	5.8715	4.4613	2.4076	2.349	2.287	2.223	2.156	2.085

Step 6: Compare calculated value (Step 2) to table value (Step 5). If calculated value > table value, reject the null hypothesis

F calculated value: 1.66 and F value from table: 2.287

1.66 < 2 .287.

So we cannot reject the null hypothesis.

Non-Parametric Test

Non-parametric test does not assume the population data belongs to some prescribed distribution which is determined by some parameters. It is also called as a **distribution-free test**. These tests are usually based on distributions that have unspecified parameters.

A non-parametric test acts as an alternative to a parametric test for mathematical models where the nature of parameters is flexible. Usually, when the assumptions of parametric tests are violated then non-parametric tests are used.

Non-Parametric Test

Def: A **non-parametric test** can be defined as a test that is used when the data under consideration does not belong to a parametrized family of distributions. When the data does not meet the requirements to perform a parametric test, a non-parametric test is used to analyze it.

Reasons to Use Non-Parametric Tests

- When the distribution is skewed. For skewed distributions, the mean is not the best measure of central tendency, hence, parametric tests cannot be used.
- If the size of the data is too small then validating the distribution of the data becomes difficult.
- If the data is nominal or ordinal, a non-parametric test is used, because a parametric test can only be used for continuous data.

Types of Non-Parametric Tests

Non parametric Tests	Parameteric Tests
Mann-Whitney U Test	Independent Samples T-tests
Wilcoxon Signed Rank Test	Paired Samples T-test
Kruskal Wallis Test	One way ANOVA
Signed Test	

Mann-Whitney U Test

The Mann-Whitney U test / Wilcoxon rank-sum test / Mann-Whitney-Wilcoxon test, is a nonparametric statistical test used to compare two independent groups of data. It is used to determine if there is a statistically significant difference between the medians of the two groups.

The test works by assigning ranks to the combined data set and then calculating the sum of the ranks for each group. The test statistic U is calculated from these sums of ranks and is compared to critical values in a table to determine statistical significance.

The null hypothesis is that there is no difference between the distributions of the two groups.

The alternative hypothesis is that there is a difference between the distributions.

It is particularly useful when the data do not meet the assumptions of normality required by parametric tests such as the t-test.

The Mann-Whitney U test only tests for a difference between the medians of the two groups and does not provide information on the direction or size of the difference. Additionally, it assumes that the two groups have similar shapes and variances, so if these assumptions are violated, alternative nonparametric tests may be more appropriate.

Null Hypothesis: H_0: The two populations under consideration must be equal.

Test Statistic: U should be smaller of:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Where R_1 is the sum of ranks in group 1 and

R_2 is the sum of ranks in group 2.

Decision Criteria:

Reject the null hypothesis if U $\leq$ critical value.

Ex: Suppose we want to compare the heights of two groups of students, group A and group B. Data is:

Group A: 165, 170, 172, 175, 178, 180, 185

Group B: 160, 165, 168, 170, 175, 180

Sol:

To perform the Mann-Whitney U test, we first combine the data from both groups and assign ranks based on the combined set of values, ignoring ties:

Combined data: 160, 165, 165, 168, 170, 170, 172, 175, 175, 178, 180, 180, 185

Ranking: 1, 2, 2, 4, 5, 5, 7, 8, 8, 10, 11, 11, 13

Next, we calculate the sum of ranks for each group:

Group A: 1 + 7 + 8 + 10 + 12 + 13 + 13 = 64

Group B: 2 + 3 + 5 + 6 + 8 + 11 = 35

We can then calculate the test statistic U, which is the smaller of the two sums of ranks:

U = min(35, 64) = 35

Then use a table of critical values to determine if the test statistic is significant. For a two-tailed test with a significance level of 0.05 and 7 degrees of freedom (n1 + n2 - 2), the critical value is 22.

Test statistic U is greater than the critical value, so we reject the null hypothesis and conclude that there is a significant difference in the median heights of the two groups.

Note that in the above example, we do not have to assume that the data are normally distributed, making the Mann-Whitney U test a useful alternative to the t-test when this assumption is violated.

Sign Test

It is a nonparametric statistical test used to compare two related or paired samples. It is used to determine if there is a statistically significant difference between the medians of the two samples.

The test works by first calculating the difference between each pair of matched observations. These differences are then categorized as positive or negative (or zero if there is no difference). The test statistic is then calculated as the number of positive differences, and the null hypothesis is that the number of positive differences is equal to the number of negative differences.

The sign test is a simple and intuitive test, and it is particularly useful when the data do not meet the assumptions of normality required by parametric tests such as the paired t-test or the Wilcoxon signed-rank test. It is also useful when the data are measured on a nominal or ordinal scale, as it does not require any assumptions about the distribution of the data.

However, the sign test has less power than other nonparametric tests, and it may not be appropriate for small sample sizes. Additionally, it only tests for a difference between the medians of the two related samples and does not provide information on the direction or size of the difference.

It is a useful alternative to parametric tests when assumptions of normality are not met or when the data are measured on a nominal or ordinal scale.

This non-parametric test is the parametric counterpart to the paired samples t-test. The sign test is similar to the Wilcoxon sign test.

Null Hypothesis: H_0: The difference in the median is 0.

Test Statistic: The smaller value among the number of positive and negative signs.

Decision Criteria: Reject the null hypothesis if the test statistic $\leq$ critical value.

Ex: Suppose we want to compare the effectiveness of two pain relief treatments, A and B, in a group of patients. We have the following data:

Patient	Treatment A	Treatment B	Difference (A - B)
1	5	7	-2
2	8	5	3
3	7	6	1
4	4	4	0
5	6	7	-1
6	5	5	0
7	6	4	2
8	7	6	1
9	6	6	0
10	5	4	1

Sol:

To perform the sign test, we first calculate the difference between the two treatments for each patient, as shown in the "Difference (A - B)" column. We then categorize each difference as positive, negative, or zero:

Patient	Treatment A	Treatment B	Difference (A - B)	Sign
1	5	7	-2	-
2	8	5	3	+
3	7	6	1	+
4	4	4	0	0
5	6	7	-1	-
6	5	5	0	0
7	6	4	2	+
8	7	6	1	+
9	6	6	0	0
10	5	4	1	+

Count the number of positive signs (in this case, 4) and the number of negative signs (in this case, 3). We can then use the binomial distribution to determine if the number of positive signs is statistically significant.

Assuming a two-tailed test with a significance level of 0.05, we need to determine the probability of observing 4 or more positive signs (or 4 or more negative signs) out of 7 total signs under the null hypothesis that the treatments are equally effective.

Using the binomial distribution, we find that the probability of observing 4 or more positive signs (or 4 or more negative signs) out of 7 total signs is approximately 0.3438. Since this probability is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no statistically significant difference between the effectiveness of treatments A and B.

Note that in this example, we only tested for a difference between the medians of the two treatments and did not provide information on the direction or size of the difference. If we want to test for the direction of the difference, we could use the one-tailed version of the sign test instead.

Kruskal Wallis Test

It is a nonparametric statistical test used to determine if there are any significant differences between three or more independent groups. It is used when the data do not meet the assumptions of normality and/or equal variances required by parametric tests such as the one-way ANOVA.

The test works by ranking the observations from all groups together, and then calculating the sum of ranks for each group. The test statistic is then calculated as:

$H = [12/(n(n+1))] * \Sigma(Rj^2/nj) - 3(n+1)$

where:

n is the total number of observations

nj is the number of observations in the j-th group

Rj is the sum of ranks for the j-th group

The null hypothesis is that there is no difference between the medians of the groups, and the alternative hypothesis is that at least one group has a different median than the others. The test statistic H follows a chi-square distribution with k-1 degrees of freedom, where k is the number of groups.

If the p-value associated with the test statistic is less than the significance level, we reject the null hypothesis and conclude that there is a significant difference between at least one pair of groups. If the p-value is greater than the significance level, we fail to reject the null hypothesis and conclude that there is no significant difference between the groups.

The Kruskal-Wallis test is useful for comparing multiple groups when the data do not meet the assumptions of parametric tests. However, it does not provide information on which specific groups are different from each other. In such cases, post-hoc tests such as the Dunn's test or the pairwise Wilcoxon rank-sum test can be used to determine which specific groups differ from each other.

The parametric one-way ANOVA test is analogous to the non-parametric Kruskal Wallis test. It is used for comparing more than two groups of data that are independent and ordinal.

Null Hypothesis: H_0: m population medians are equal

Test Statistic: $H = \left(\frac{12}{N(n+1)} \sum_1^m \frac{R_i^2}{R_j}\right) - 3(N+1)$

where, N = total sample size, n_j and R_j are the sample size and the sum of ranks of the jth group

Decision Criteria: Reject the null hypothesis if H $\geq$ critical value

Ex: Suppose patients are suffering from chikungunya. They are divided into three groups and different drugs were administered. The platelet count for the patients is given in the table below. It needs to be checked if the population medians are equal. The significance level is 0.05.

Drug 1	Drug 2	Drug 3
42000	67000	78000
48000	57000	89000
57000	79000	67000
69000		80000
45000		

Sol:

As the size of the 3 groups is not same the Kruskal Wallis test is used.

H_0: Population medians are same

H_1: Population medians are different

$$n_1 = 5, n_2 = 3, n_3 = 4$$

N = 5 + 3 + 4 = 12

Now ordering the groups and assigning ranks:

Drug			Rank		
1	2	3	1	2	3
42000			1		
45000			2		
48000			3		
57000	57000		4.5	4.5	
	67000	67000		6.5	6.5
69000			8		
		78000			9
	79000			10	
		80000			11
		89000			12

$$R_1 = 18.5, R_2 = 21, R_3 = 38.5$$

Substituting these values in the test statistic formula:

$$\left(\frac{12}{N(n+1)}\sum_{1}^{m}\frac{R_i^2}{R_j}\right) - 3(N+1)$$

H = 6.0778.

Using critical value table, critical value will be 5.656.

As H < critical value, the null hypothesis is rejected and it is concluded that there is no significant evidence to show that the population medians are equal.

Difference between Parametric and Non-Parametric Test

Non-Parametric Test	**Parametric test**
It is used when the population data does not belong to a parametrized distribution.	It is used when the data belongs to a specific probability distribution such as a normal distribution.
Knowledge of the population is not required to conduct this test.	Complete knowledge of the population is required.
The central tendency value used is the median .	mean is used
It is used for ordinal data and nominal data.	It is used for interval data.
Less powerful	More powerful t
Examples of non-parametric tests are signed test, Kruskal Wallis test, etc.	Examples of parametric tests are z test, t test, etc.

Advantages of Non-Parametric Test

- Knowledge of the population distribution is not required.
- The calculations involved in such a test are shorter.
- A non-parametric test is easy to understand.
- These tests are applicable to all data types.

Disadvantages of Non-Parametric Test

- They are not as efficient as their parametric counterparts.
- As these are distribution-free tests the level of accuracy is reduced.

Ex 1: A surprise quiz was taken and the scores of 6 students are given as follows:

Student	1	2	3	4	5	6
Score	8	6	4	2	5	6

The same quiz was taken again after one month, and the following scores were obtained.

Student	1	2	3	4	5	6
Score	6	8	8	9	4	10

Using a non-parametric test identify if there is a difference in the marks obtained. The significance level is 0.05.

Solution:

The Wilcoxon signed rank test will be used.

Student	Test 1 Score	Test 2 Score	Difference (Test 2 - Test 1)
1	8	6	-2
2	6	8	2
3	4	8	4
4	2	9	7
5	5	4	-1
6	6	10	4

Assigning signed ranks to the differences

Difference	Rank	Signed Rank
-1	1	-1
2	2.5	2.5
-2	2.5	-2.5
4	4.5	4.5
4	4.5	4.5
7	6	6

H_0: Median difference is 0.
H_1: Median difference is positive.
W1: Sum of positive ranks = 17.5
W2: Sum of negative ranks = 3.5
As W2 < W1, thus, W2 is the test statistic.
Now from the table, the critical value is 2.
Since W2 > 2, thus, the null hypothesis cannot be rejected and it can be concluded that there is no difference between the scores of the two tests.

Answer: Fail to reject the null hypothesis

Ex 2: Use the sign test to solve example 1.

Solution:

Student	Test 1 Score	Test 2 Score	Difference (Test 2 - Test 1)	Sign
1	8	6	-2	-
2	6	8	2	+
3	4	8	4	+
4	2	9	7	+
5	5	4	-1	-
6	6	10	4	+

H_0: Median difference is 0.
H_1: Median difference is positive.
Number of (-) signs = 2
Number of (+) signs = 4
As number of (-) signs < number of (+) signs, thus, the test statistic = 2
Now from the table, the critical value is 6.
As 2 < critical value, thus, the null hypothesis is rejected and there is no evidence to suggest that the median difference is 0.
Answer: Null hypothesis is rejected

Example 3: A test was run on 5 patients to see if a new drug could cure sleepwalking. Another group of 5 patients was still taking the old drug. The number of sleepwalking cases in a month is as follows:

Sleepwalking cases in a month	New Drug	7	8	4	9	8
	Old Drug	3	4	2	1	1

Using a non-parametric test check if there is a difference in the number of sleepwalking cases. The significance level is 0.05.

Solution:

The Mann Whitney U test is used. Ordering the data and assigning ranks

Drug		Rank	
Old	New	Old	New
1		1.5	
1		1.5	
2		3	
3		4	
4	4	5.5	5.5
	7		7
	8		8.5
	8		8.5
	9		10

H_0: Two groups report same number of cases
H_1: Two groups report different number of cases
R_1= 15.5, R_2 = 39.5

$$n_1 = n_2 = 5$$

Using the formulas,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 24.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 0.5$$

As $U_2 < U_1$, thus, U_2 is the test statistic.
From the table the critical value is 2
As U_2 < 2, the null hypothesis is rejected and it is concluded that there is no evidence to prove that the two groups have the same number of sleepwalking cases.
Answer: Null hypothesis is rejected

Thanks

Any questions ?